EDUCATIONAL MEASUREMENT

IN THE

ELEMENTARY GRADES

BY I. N. MADSEN, PH.D.

Director of the Department of
Tests and Measurements, State Normal School
Lewiston, Idaho

WORLD BOOK COMPANY

THE HOUSE OF APPLIED KNOWLEDGE

"UNTO one he gave five talents, to another two, to another one; to every man according to his several ability." As old as Scripture is recognition of the fact that capacities vary, but new as our generation is the application of scientific method to the determination of individual differences. Much has been done in the way of providing tests for students in the elementary grades of our schools, and many tests for making various classifications are now available; but too often teachers do not understand the theory underlying the tests that they use; too often they do not know how to get the maximum benefit from testing. For the guidance of teachers who wish to be informed on the technique of testing in order that they may render the best possible service to their charges, this book on *Educational Measurement in the Elementary Grades* has been prepared.

Copyright 1930 by World Book Company
Copyright in Great Britain
All rights reserved

PRINTED IN U.S.A.

PREFACE

THERE are now many excellent books available in the field of educational measurements. None, however, deals comprehensively and solely with all the major phases of testing for the elementary grades. In writing this book the author has had in mind the training of teachers, supervisors, principals, and others whose chief concern will be with these grades. It is here that we find a great majority of the school children, as well as the greatest development of standardized tests. Yet discouragingly little use is made of such tests for the improvement of teaching. A splendid tool lies idle or is unskillfully used because of the inadequacy of the training given craftsmen.

Many normal schools and teachers' colleges are attempting to remedy this situation by including in their curricula required courses in educational tests and measurements. But it is usually not possible to devote to these more than twelve weeks, with three or four hours of instruction a week. Most students in these institutions are certified to teach after two years, or less, of training. This means that instruction in educational measurements, as in other subjects, must be limited to the minimum essentials. An attempt has therefore been made to include in this book only topics which have been found most significant for the elementary grades.

During ten years of experience in teaching beginning students, the author has become convinced that the best results are obtained by pointing out, as definitely and concretely as possible, the relation of testing to such other important phases of teaching as the course of study, the determination of objectives and their attainment, pupil diagnosis, remedial instruction, sectioning classes, classifying pupils, etc. Wherever possible, this relationship has been pointed out. Fortunately an abundance of laboratory material is available to supplement the theoretical work in a measurements course. Thus the students can participate in giving

the tests, they can score the results, tabulate or otherwise treat results to provide a basis for their interpretation, etc.

A discussion of statistical concepts has been placed in the first part of the book, since a relatively early mastery of these concepts will greatly facilitate mastery of the later chapters. The instructor should find no difficulty in obtaining local material to demonstrate the usefulness of statistical concepts and to provide the necessary drill or practice in applying them. It is hoped that Chapters V and VI may give the student an appreciation of the importance of the relation between intelligence and progress in school and that they may also make possible more intelligent participation in testing programs conducted by experts. It may be said that many teachers misinterpret intelligence tests; however, the same can be said about achievement tests and, for that matter, about any other method of pupil investigation. Should valuable but complicated tools be discarded because of clumsy workmen? Is it not more logical to insist on better training and higher standards? In selecting achievement tests for discussion in Chapters VII and VIII, the aim was to get a cross-section of available tests of this type rather than to apply rigid criteria of merit. While there is no doubt that great progress has been made in the evaluating of tests, we have not yet reached the point where the doctors agree.

The author is indebted to many who have helped directly or indirectly in the preparation of this book. He is particularly indebted to Dr. Lewis M. Terman for many helpful suggestions. The inspiration that he has received from the work of all the contributors to the theory and practice of educational measurement is as gratefully acknowledged; wherever possible, specific acknowledgment has been made in the text to individuals of this group who have been directly or indirectly quoted.

I. N. MADSEN

CONTENTS

v

EDITOR'S INTRODUCTION

THIS book by Dr. Madsen, *Educational Measurement in the Elementary Grades*, fills a long-existing gap in the *Measurement and Adjustment Series*. On a subject which is so new and which has so many technical aspects it is not easy to write a book suitable for the average teacher or for the beginning student in normal schools and teachers' colleges. In attempting to prepare such a book the writer is faced by two dangers. If he writes a book that is readily comprehensible, he is likely to omit or slight important technical considerations without an understanding of which the results of educational measurements cannot be properly interpreted. If on the other hand he is careful to avoid this danger, the result is likely to be a book better adapted to the graduate student than to the undergraduate or the teacher-in-service.

Of the half-dozen or more manuscripts which have been submitted to this series as strictly introductory texts in educational measurement, this by Dr. Madsen is the first that the Editor has been willing to recommend for publication. The author himself would be the last to claim that this book is free from faults, but it can fairly be said that it provides an excellent orientation to the student who is entering upon the subject for the first time. In the opinion of the Editor, Dr. Madsen's book gives about all the information on this subject that can reasonably be regarded as essential for the rank and file of teachers to have. It is rather generally agreed that all who are preparing to teach should have at least one general course dealing specifically with educational tests as distinct from intelligence tests. However, to attempt to make every teacher an expert in the use and interpretation of educational measurements would be a mistake; there are too many other things which it is important for the teacher

to get, not the least of which is a fair degree of expertness in the teaching process.

As is clearly indicated by its title, this book has been prepared exclusively for those who teach or expect to teach in the elementary grades. There is no need of burdening the elementary teacher with information about measurement methods designed for use in the high school or college grades. Those fields have been admirably dealt with by two earlier texts in this series; namely, *Tests and Measurements in High School Instruction*, by Ruch and Stoddard, and *Measurement in Higher Education*, by Wood. A splendid critical treatment of the whole subject of educational testing, for the more advanced student, will be found in Kelley's *Interpretation of Educational Measurements*, also in this series.

The content and organization of this book have gradually taken form over a period of several years in teaching the subject to students in a normal school. The book contains no material which has not been extensively tried out with students of the academic level for which it is designed. It is this fact, doubtless, which is responsible for the author's simple style and straightforward exposition.

LEWIS M. TERMAN

EDUCATIONAL MEASUREMENT IN THE ELEMENTARY GRADES

CHAPTER ONE

The Measurement of Individual Differences

Origin and development of the testing movement. Individual differences in human beings have probably always been noted and remarked upon. However, it is only during recent times that scientific observations and measurements of such differences have been made, and it is even more recently that the measurement of these differences has been made of practical use in education. In 1879 Wundt established the first laboratory for the scientific study of psychology. A few years later, in 1884, Sir Francis Galton founded his laboratory for anthropometric measurements. Long before this, however, Galton had been interested in the study of the heredity of mental traits and capacities, and in 1869 had published his *Hereditary Genius*. The testing movement was definitely introduced into America as early as 1890 by an American psychologist, J. McKeen Cattell, who had conducted experiments in Wundt's laboratory and who had later been associated with Galton. At this time Cattell outlined a testing program which was published in the English journal, *Mind*, and during this period he also began experimenting with tests for the measurement of mental traits.

For the fifteen years following 1890 psychologists were busily experimenting with tests that would effectively describe human intelligence. It was not until 1905, however, that Alfred Binet published his first rough scale; this he standardized on an age basis in 1908, and revised in 1911.

Coincident with these experiments in the measurement of mental traits, we find a movement which concerned itself with the progress and achievement of pupils in school. One phase of this movement dealt with the age-grade status of pupils and their elimination from school. Among the best-known investigations of this kind are those by Thorndike,[1] Ayres,[2] and Strayer,[3] published 1907, 1909, and 1911 respectively. These studies not only aroused widespread interest and discussion but resulted in a flood of similar investigations. In general, the purpose of such investigation was to ascertain the facts concerning age-grade status and elimination, to use these facts in determining the efficiency of the schools, to study causes for the conditions found, and to fix responsibility. The most striking fact revealed by these studies was the wide range of individual differences in the progress of pupils through school.

The second phase of the movement dealt with the development and the use of standardized tests. Dr. J. M. Rice,[4] as a result of his investigation of the efficiency in teaching spelling, is usually given credit for being the first to use comparative tests for measuring the results of teaching. Rice's investigation began in 1894 and attracted much attention for several years thereafter. During the year 1908 C. W. Stone,[5] under the direction of Professor Thorndike, developed a

[1] E. L. Thorndike, *The Elimination of Pupils from School* (United States Bureau of Education Bulletin, No. 4). Government Printing Office, Washington; 1907.

[2] Leonard P. Ayres, *Laggards in Our Schools.* Charities Publication Committee, New York; 1909. (Now published by Survey Associates, Inc., 112 East Nineteenth Street, New York.)

[3] George D. Strayer, *Age and Grade Census of Schools and Colleges* (United States Bureau of Education Bulletin, No. 5). Government Printing Office, Washington; 1911.

[4] J. M. Rice, "The Futility of the Spelling Grind," in *The Forum,* Vol. XXIII, pages 163–172, 409–419; 1897.

[5] C. W. Stone, *Arithmetical Abilities and Some Factors Determining Them* (Contributions to Education, No. 19). Teachers College, Columbia University, New York; 1908.

standardized arithmetic test; and in the course of the following year Thorndike's [1] handwriting scale appeared. It is apparent, therefore, that scientific methods for measuring the results of teaching, like those for measuring intelligence, are of recent origin.

The scientific measurement of human traits may conveniently be summarized under five heads: (1) anthropometric measurements, which are concerned with the measurement of different parts of the human body, such as height, weight, head girth, etc.; (2) measurements of sensory acuity, which measure the keenness of the sense organs — visual, auditory, tactual, etc.; (3) measurements of reaction time and motor ability, such as the speed of reaction to visual, auditory, or other sensory stimuli; (4) measurements of complex mental processes, such as memory, imagery, etc.; (5) measurements of general mental ability, special aptitudes, achievement in school, etc. It is this last group that has been found most significant in teaching, and consequently it will be our primary concern in this book.

The nature of individual differences. All the traits that are listed above, when measured, show wide individual differences. The nature of these differences can be best illustrated by arranging the measurement of a given trait in the form of a frequency table or curve. The following tables and curves are based upon actual measurement of the traits specified. The first table shows the range and distribution of height in a group of 202 women students in normal school. The same facts are represented graphically in Figure 1.

From Table 1 it will be seen that the 202 women composing the group range in height from 57 to 69 inches, that the mean (average) is 63 inches, that the greatest proportion of measurements cluster about the mean, and that the frequencies

---

[1] E. L. Thorndike, "A Handwriting Scale," in *Teachers College Record,* Vol. II; March, 1910.

TABLE 1

<small>Frequency Table Showing the Range and Distribution of Height in a Group of Women Students</small>

Height in Inches	Frequency
69	1
68	3
67	8
66	19
65	26
64	32
63	36
62	29
61	24
60	15
59	6
58	2
57	1
Total	202
Mean	63

decrease in about the same way on either side of the mean. Results corresponding very closely to these would be obtained in any similar group of women. Indeed, accurate measurement of any one trait in such a group would reveal the same characteristic tendencies. These tendencies may be generalized as follows. (1) The abilities are continuous. They range without a break from the lowest to the highest. In Table 1 there appears to be a break of one inch between each group of frequencies and the group above or below. However, this is not a true break, since any one group of measures, such as the fifteen cases given opposite the 60 interval, include all the measures that range from 60 up to, but not including, 61. Sometimes breaks may appear to exist when a small
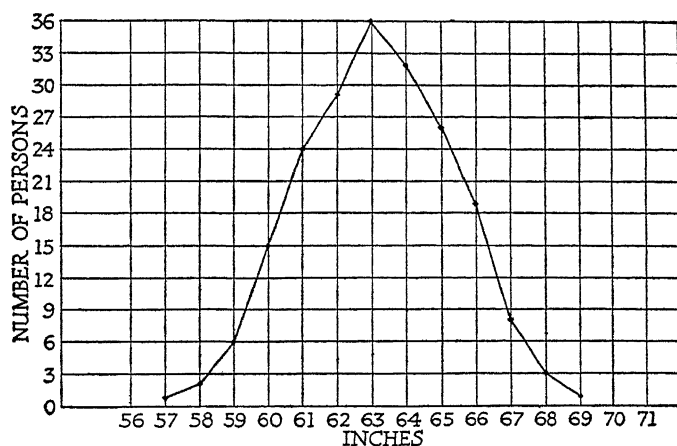
FIG. 1. Distribution of the heights of 202 women.
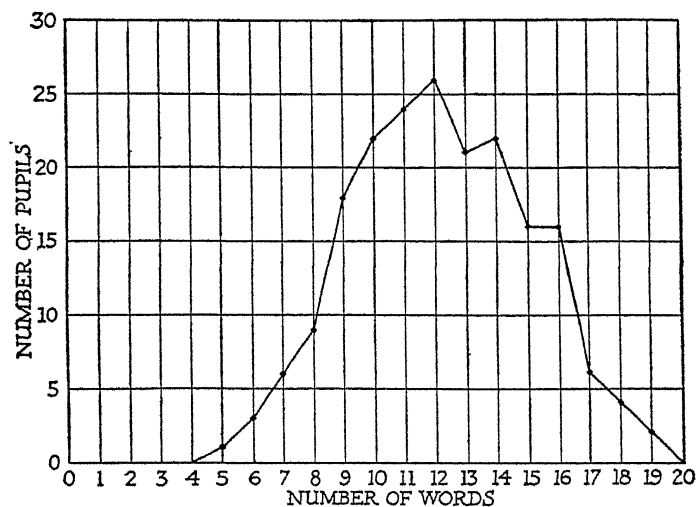
FIG. 2. Distribution of words correctly spelled by 186 fifth-grade pupils.

group is measured, but they are always filled in when additional measurements are made. (2) The abilities tend to cluster about the center of the distribution in such a way that a large proportion of measures are fairly close to the mean in magnitude. Thus, referring again to Table 1, we see that 97 women, or approximately one half of the total number, range in height from 62 to 64 inches inclusive. (3) The measures decrease in about the same proportion on both sides of the mean. When a frequency table is plotted in the form of a frequency curve, the result shows a tendency toward a symmetrical, bell-shaped curve. (4) Variations in a trait appear to be distributed in accordance with the law of chance as illustrated by coin tossing. This is significant in connection with the mathematical properties of the frequency curve
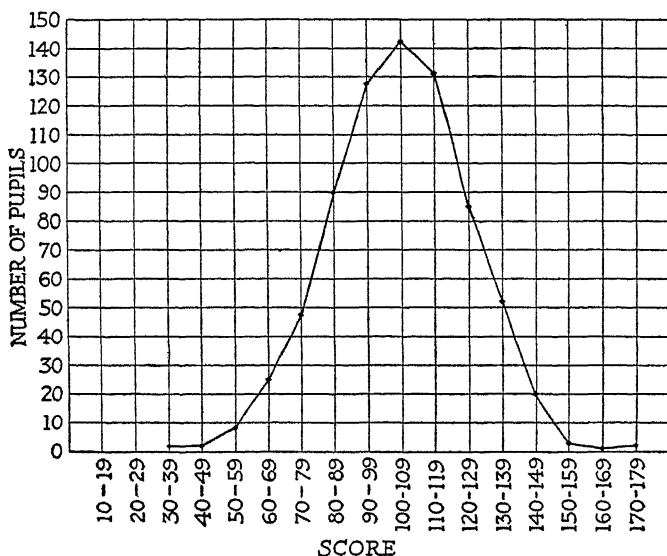
Fɪɢ. 3.   Distribution of scores in Haggerty Intelligence Examination, Delta 2, by 737 seventh-grade pupils.
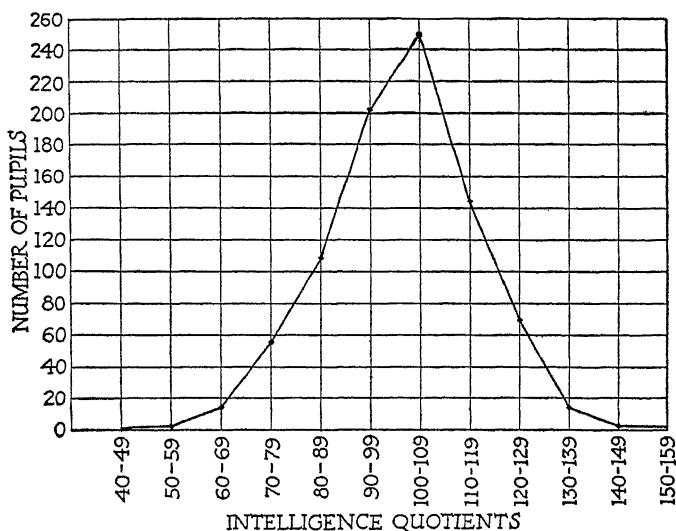
FIG. 4. Distribution of IQ's of 880 pupils according to the Stanford-Binet Test.[1]

and will be commented upon elsewhere. It should be emphasized that the characteristics listed above tend to appear only when the measurements are made accurately and with sufficient numbers. These tendencies are more concretely illustrated in the curves shown in Figures 1 to 4.

Importance of individual differences in school progress. It is obvious that individual differences, such as those represented in Figures 2 to 4, are of great importance in determining the progress of pupils through school. Largely because of these differences we find that pupils vary widely in their rate of progress and in the extent to which they profit from instruction. Starch quotes statistics from the St. Louis schools that are of interest in this connection. For many years these schools had promoted pupils at the end of each

[1] Adapted from I. N. Madsen, "Some Results with the Stanford Revision of the Binet-Simon Tests," in *School and Society*, May 10, 1924.

TABLE 2

THE AVERAGE ATTENDANCE PER GRADE OF 1439 PUPILS, GRADUATES, RE-
QUIRED TO COMPLETE EACH OF THE EIGHT GRADES.   FORTY WEEKS IS
ASSIGNED IN THE COURSE OF STUDY FOR EACH GRADE.*

AVERAGE NUMBER OF WEEKS TO EACH GRADE	NUMBER OF PUPILS	AVERAGE NUMBER OF WEEKS TO EACH GRADE	NUMBER OF PUPILS
17	1	44	27
18	2	45	19
19	0	46	20
20	1	47	15
21	1	48	9
22	8	49	5
23	8	50	4
24	13	51	4
25	17	52	2
26	19	53	2
27	25	54	2
28	46	55	1
29	43	56	2
30	52	57	2
31	83	58	2
32	103	59	1
33	99	60	1
34	109	61	0
35	92	62	2
36	110	63	1
37	87	64	0
38	104	65	0
39	95	66	0
40	87	67	0
41	33	68	0
42	49	69	0
43	29	70	2

Median   . . . .   35 weeks
Total average time to do 320 weeks' work .   288 weeks
Double promotions 17%
Normal promotions 67%
Repeaters 16%

ten weeks, according to the efficiency of their work. Thus some pupils advanced rapidly, some at a medium rate, and some slowly. Table 2 shows the variation in the time required by 1439 pupils to complete the work of the eight elementary grades. Reading this table from the top, we see that one pupil required an average of 17 weeks to complete each grade of forty weeks' work, two pupils required 18 weeks, one pupil 20 weeks, another 21 weeks, and so on. At the other extreme we find two pupils who require 70 weeks to complete each grade of forty weeks' work. In other words, some pupils can do the work of a grade in approximately half a year, while others require nearly two years to a grade.

Starch also quotes statistics from Thorndike showing " the range of ages of boys and girls in the third year of high schools in Chicago, Philadelphia, New York, Detroit, Fall River, Los Angeles, Lowell, and Worcester," as follows:

TABLE 3

SHOWING VARIATIONS IN AGES OF BOYS AND GIRLS IN THIRD YEAR OF
HIGH SCHOOLS *

AGE	13	14	15	16	17	18	19	20 OR MORE	TOTAL
Boys	7	92	594	1246	1203	572	193	67	3974
Girls	4	73	562	1351	1289	554	120	34	3987
Total	11	165	1156	2597	2492	1126	313	101	7961

*Adapted from Starch's *Educational Psychology*, page 79. By permission of the Macmillan Company, publishers.

The figures in Table 3 were collected originally to show sex differences in variability. However, they also show for both sexes a wide range in the ages at which pupils reach the third year of high school. These figures corroborate the results

shown in Table 2 in that they also indicate differences in the rate of progress through school. Similar data may be obtained in any typical school system by tabulating in the form of an " age-grade " table the variations in age that occur in each grade. This method is illustrated in Table 4. The data for this table were obtained in connection with the testing of pupils in Grades III to XII inclusive and are typical of the age-grade distribution in most schools. The table shows strikingly the wide range in ages for each grade. For example, in Grade III pupils range in age from 7 to 16, in Grade IV from 8 to 14, in Grade V from 8 to 16, and so on. Assuming that the normal age for entering Grade I is 6 or 7, a range of two years is allowed in each grade for the normal or " at-age " group. The accelerated pupils are those young for their grade, and the retarded are those old for their grade. At the bottom of the table are shown the number and the percentage of each of these three groups in each grade.

The determination of the causes of differences in school progress. It can readily be shown that the differences indicated in Tables 2, 3, and 4 are due largely to differences in intelligence. In Table 5 the mean mental age has been computed for each group in Grades I to VIII. This table shows that, while the pupils in each grade vary greatly in chronological age, the variation in median mental age is small. For example, in Grade I the variation is only 1.3 years, in Grade II it is 1.9 years, etc. In other words, an important cause of the great difference in chronological age in a given grade is the difference in the rate of mental growth. In later chapters it will be shown that factors such as faulty teaching in the various school subjects may also retard the progress of pupils. Whatever the causes may be, it is clear that they can best be discovered by the use of suitable standardized tests. In later chapters we shall discuss tests that are available for this purpose.

TABLE 4*

SHOWING AGE-GRADE DISTRIBUTION OF PUPILS IN GRADES III TO XII

AGE	3	4	5	6	7	8	9	10	11	12	TOTALS
7	10										10
8	242	48	2								292
9	151	323	58	1							533
10	79	303	331	55							768
11	27	164	323	260	36	5					815
12	15	75	173	286	196	42	8	1			796
13	1	19	85	193	242	221	155	7			923
14	1	16	33	83	162	222	791	127	8		1443
15			12	30	83	133	1126	589	117	9	2099
16	1		4	6	16	57	527	689	538	87	1925
17				2	2	10	163	332	523	490	1522
18							31	69	204	404	708
19							12	19	63	147	241
20							2	10	11	50	73
21									3	17	20
22										3	3
Totals	527	948	1021	916	737	690	2815	1843	1467	1207	12171
Accelerated	10	48	60	56	36	47	163	135	125	96	776
At-age	393	626	654	546	438	443	1917	1278	1061	894	8250
Retarded	124	274	307	314	263	200	735	430	281	217	3145
% Accelerated	2	5	6	6	5	7	6	7	8	8	6
% At-age	75	66	63	60	60	65	68	70	72	74	68
% Retarded	23	29	31	34	35	28	26	23	20	18	26

*I. N. Madsen, "Intelligence as a Factor in School Progress," in *School and Society*, Vol. XV, pages 283–288; March 11, 1922.

11

TABLE 5

SHOWING MEDIAN MENTAL AGES BY GRADES AND CHRONOLOGICAL
AGES *

GRADE	CHRONOLOGICAL AGES												NUMBER OF PUPILS
	5	6	7	8	9	10	11	12	13	14	15	16	
I	6.6	6.8	7.3	7.2	7.3	7 3	6 0						536
II		7.6	7.8	8 2	8 3	8 0	9.5	8.5					166
III				8.7	8 6	8 5	8 5						499
IV				9.8	10 2	10 1	9.6	9.6	9 6				932
V					11.7	11.7	11 1	10 9	10 8	10.0			1004
VI						14.1	13 3	12.7	12 6	12.6	11.0		907
VII							14.9	14 6	14 2	13.6	13.2	12.5	736
VIII								15 8	15 5	15.0	14 1	14.5	675

* In this table the mental ages in Grades I and II were derived from the Stanford-Binet Tests, while those in Grades III to VIII were derived from the Haggerty Intelligence Examination, Delta 2. The table is adapted from: I. N. Madsen, "Some Results and Uses of Intelligence Tests in the Schools of Idaho," in *Lewiston Normal School Bulletin*, Vol. XV, pages 4, 20, and 21; April, 1924.

The table is read as follows : In Grade I the mean mental age of the five-year-olds is 6.6 years, of the six-year-olds 6.8 years, of the 7-year-olds 7.3 years, etc.

The need of differential treatment of pupils. It is clear from the foregoing discussion that because pupils differ from each other in so many ways, they will not fit equally well into any program or course of studies. The consequences and implications of this fact will be discussed in detail in later chapters. It will suffice at this time to suggest briefly some problems of teaching and of curriculum organization which arise in connection with the discovery of important individual differences among pupils.

First, it would be unreasonable to expect all of a group of pupils in a given grade to complete a unit of assigned work in the same specified time. We have seen from Table 3 that some pupils can work three or four times as rapidly as others. It follows that if we insist that all of them do the same work,

we must give some of them more time than others. This, of course, would necessitate varying promotional rates for pupils who have different capacities.

Second, if we insist on keeping pupils of approximately the same age together through the grades — that is, require the same time allotment for all — it becomes necessary to vary the amount of work done by each pupil in a unit of time. This situation gives rise to the problem of differentiating subject matter according to the differences among pupils. That is, if we are to give adequate recognition to individual differences, it may become necessary to give some pupils only the minimum essentials of a subject, while for others a richer content is required.

Third, it may become necessary in the higher grades to provide a large number of subjects from which the pupils may choose electives best suited to their needs.

This leads us to a fourth problem ; namely, educational and vocational guidance. Because the function of the school is to do what it can to fit each child for life, it becomes necessary to take stock of the various capacities and achievements of each child from time to time, and, on the basis of the findings, to give him the training that will develop his latent capacities to the best advantage for himself and for society.

EXERCISES

1. Plot a frequency curve from the data in Table 2.

2. Plot two frequency curves from the data in Table 3, one for the boys and one for the girls.

3. Make an age-grade table from data obtained from your local schools.

4. Obtain accurate measurements of the heights of a group of women students, not less than 30 in number, and tabulate in the form of a frequency table, such as Table 1. Obtain similar measurements for a group of men students and tabulate in the same manner.

5. Do the two tables obtained from Exercise 4 manifest the tend-

encies described on pages 3 to 7? Attempt to account for any deviations from these tendencies.

6. Obtain the final grades for a class of 30 or more students in some subject and tabulate in the form of a frequency table. Compare the results obtained with those obtained in Exercise 4.

7. Do you think that the measurement of a student's achievement in terms of "grades" is as trustworthy as the measurement of such physical traits as height and weight? Explain.

CHAPTER TWO

The need for measurement in teaching. Thorndike has
defined education as the " making of useful changes in human
beings." Educators may differ as to what constitute *useful*
changes, but they can hardly take exception to this definition
as a general proposition. This being the case, measurement
becomes necessary from time to time in order to determine
the changes brought about as the result of teaching. A few
years ago the author wrote :

Testing is and indeed always has been considered an indispensable
aid to teaching. Teachers have always endeavored to measure
progress of pupils towards a goal and to diagnose defects by means
of testing. The development and use of standard tests may there-
fore be regarded as the extension and improvement of an old device.
They are more precise and exact than ordinary teachers' exam-
inations and so accomplish the purpose of testing more reliably.
They enable us to set up definite goals of achievement because
measurements are more objective and less influenced by personal
judgment. They enable us to set up natural and attainable norms
or standards of achievement for any given grade because they are
based upon the actual attainments of pupils under typical school
conditions.[1]

To be sure, there are those who profess to believe that any
testing or examining of pupils does more harm than good.
The author recalls a lecturer who argued that pupils do much
better work when no tests or examinations are given. The
question immediately arises : How can one be sure of this
without testing ?

Objective versus subjective measurement. Two general
methods are available in the measurement of amount or
quantity. They have been called the *objective* and the *sub-*

[1] I. N. Madsen, *A Teachers' Guide for the Use of Standard Tests* (test
bulletin published and copyrighted by the author, 1925), page 1.

jective methods. It is the subjective method that is used when one estimates a person's weight, or the temperature, or the distance between two points. When this first method is used, different estimators are likely to differ greatly in their guesses. Each guess has a personal or subjective reference. The second, or objective, method is used when one applies a definite instrument of measurement, such as a yardstick for measuring distance, a standard scale for measuring weight, or a thermometer for measuring temperature. When this method is used, different measurers will obtain nearly the same results, provided they exercise care and skill in the use of their instruments.

Everyone is familiar with objective methods of measuring amount or quantity in the physical world. So accustomed are we to the use of yards, feet, inches, pounds, degrees of temperature, etc., that we have come to take them for granted. However, these measuring devices did not always exist. For example, as the name would suggest, our present standard twelve-inch foot evolved from the use of the human foot as a unit of measurement. Since all human feet are not twelve inches in length, we can readily see that in measuring a given distance different people would not obtain the same result. It is not too much to say that progress in science is largely the result of the increase in objectivity of measurement and standardization of measuring devices, which permits greater precision and facility in measurement. Lord Kelvin, the great British scientist, is quoted as saying:

> When you can express what you are talking about in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind. . . .

Until we can really measure educational progress — express it in numbers, as Lord Kelvin says — our knowledge is likely to be of a " meager and unsatisfactory kind."

The unreliability of school marks. As suggested above, teachers have long attempted to measure the progress of pupils in their school work by giving examinations and stating the results in the form of marks or grades. Repeated investigations have shown that such marks are unreliable. A much-quoted investigation by Starch [1] showed extreme inability on the part of experienced teachers to agree on the marks to be assigned to test papers in English, mathematics, and American history. Thus the marks assigned by 142 English teachers on one English paper ranged from 64 to 98, on a percentage basis; on another English paper they ranged from 50 to 98; on a geometry paper the marks assigned by 114 teachers ranged from 28 to 92; and on an American history paper the marks assigned by 70 teachers ranged from 42 to 90.

Ashbaugh [2] reports a similar experiment in which a class of advanced university students marked a seventh-grade pupil's arithmetic paper. The problems were taken from the Stone Reasoning Test. However, the students were given neither directions nor rules for scoring, but were allowed to grade the paper as they would an ordinary classroom test. Ashbaugh required the students to grade this paper three successive times, with an interval of four weeks between the first and the second grading and a corresponding interval between the second and third gradings. The results showed two things: First, the students did not agree with each other in assigning marks to the paper in question; second, they did not agree with their own marks when making successive ratings. The questions and the pupil's paper used by Ashbaugh follow: [3]

[1] Daniel Starch, *Educational Psychology*, pages 519–521. The Macmillan Company, New York; 1927.

[2] E. J. Ashbaugh, "Reducing the Variability in Teachers' Marks," in *Journal of Educational Research*, Vol. IX, pages 185–198; March, 1924.

[3] *Op. cit.* Used by permission of the Public School Publishing Company, Publishers.

1. If you buy 2 tablets at 7 cents and a book for 65 cents, how much change should you receive from a two-dollar bill?

2. John sold 4 *Saturday Evening Posts* at 5 cents each. He kept $\frac{1}{2}$ of the money and with the other $\frac{1}{2}$ he bought Sunday papers at 2 cents each. How many did he buy?

3. If James had 4 times as much money as George, he would have $16. How much money had George?

4. How many pencils can you buy for 50 cents at the rate of 2 for 5 cents?

5. The uniforms for a baseball nine cost $2.50 each. The shoes cost $2.00 a pair. What was the total cost of uniforms and shoes for the nine?

6. In the schools of a certain city there are 2200 pupils; $\frac{1}{2}$ are in the primary grades, $\frac{1}{4}$ in the grammar grades, $\frac{1}{8}$ in the high school, and the rest in the night school. How many pupils are there in the night school?

7. If $3\frac{1}{2}$ tons of coal cost $21, what will $5\frac{1}{2}$ tons cost?

8. A news dealer bought some magazines for $1.00. He sold them for $1.20, gaining 5 cents on each magazine. How many magazines were there?

9. A girl spent $\frac{1}{8}$ of her money for carfare, and three times as much for clothes. Half of what was left was 80 cents. How much money did she have at first?

10. Two girls receive $2.10 for making buttonholes. One makes 42, the other 28. How shall they divide the money?

A Pupil's Paper on Stone's Standardized Reasoning Test
in Arithmetic

1.
$$\begin{array}{r} 7 \\ 2 \\ \hline 14 \end{array} \qquad \begin{array}{r} 14 \\ 65 \\ \hline 89 \end{array} \qquad \begin{array}{r} 200 \\ 89 \\ \hline \$1.21 \end{array}$$

2.
$$\begin{array}{r} 5¢ \\ 4 \\ \hline 2)\overline{20} \\ 10 \end{array} \qquad \begin{array}{r} 2)\overline{10¢} \\ 5 \text{ papers} \end{array}$$

3.
$$\begin{array}{r} 4)\overline{\$16} \\ 4 \end{array}$$

4. 2 pencils $= 5¢$
    1 " $= \frac{1}{2}$ of $5¢ = 2\frac{1}{2}¢$
    $50¢ \div 2\frac{1}{2}¢ = 20$ pencils

5.   2.50 $2   6.   2)2200 $\frac{3}{8}$

       18.00   9   4)1100

       $20.50 $18   8) 275

                                        34$\frac{3}{8}$

$$\text{7.} \quad \frac{\overset{3}{\cancel{21}} \text{ tons} \times \overset{1.05}{\cancel{\$5.25}}}{\underset{.7}{\cancel{3.5} \text{ tons}}} = 30.50$$

8.   5)$1.00   5)1.20   24

         20   24   20

                                   4 magazines

9. $\frac{1}{8} + \frac{3}{8} = \frac{4}{8}$

      $\frac{4}{8} = 80$ $\frac{1}{8} = 20$ $\frac{3}{8} = 1.60$

10. $\frac{28}{70} \times 2.10 = 84¢$

       $210 - 84 = 126¢$

The writer repeated Ashbaugh's experiment with five different groups of normal-school students. More than half of these students had previously had some teaching experience. They were first given the ten problems and asked to solve them. This was done to familiarize them with the nature of the problems before they scored the sample paper reproduced above. They were then asked to score this paper on a percentage basis, and the scored papers were collected without comment. Four or five weeks later the same students were given the same paper with the same request. Two ratings were thus obtained from each student on the same pupil's paper. Table 6 shows the variation in the marks between the first and second grading. Thus in the first group Scorer 1 assigned a mark of 67 on the first rating, and a mark of 75 on the second rating, while Scorer 2 assigned a mark of 68 on both ratings, etc. This table also shows considerable variation in the mean values assigned to the same paper by the different groups. Thus Group I assigned a mean value of 58.7 in the first trial and a mean value of 60.8 in the second trial, etc.

TABLE 6

SCORER	Group I		Group II		Group III		Group IV		Group V	
	1st trial	2d trial	1st trial	2d trial	1st trial	2d trial	1st trial	2d trial	1st trial	2d trial
1	67	75	60	56	$67\frac{1}{2}$	65	61	63	79	46
2	68	68	60	53	80	70	60	60	59	61
3	64	71	76	60	73	65	85	81	55	$52\frac{1}{2}$
4	56	65	51	50	62	62	70	65	71	65
5	73	70	60	50	58	45	71	71	68	64
6	56	46	60	50	54	54	80	66	58	58
7	$61\frac{1}{2}$	85	50	60	40	68	54	55	65	65
8	62	60	55	57	60	64	71	53	60	60
9	63	58	60	55	61	64	60	50	71	56
10	$41\frac{1}{4}$	$71\frac{1}{4}$	$89\frac{1}{4}$	56	65	66	50	64	66	63
11	57	60	64	60	62	58	68	53	54	53
12	58	$67\frac{1}{2}$	31	58	58	60	60	49	50	56
13	43	51	$63\frac{1}{4}$	54	55	61	54	45	51	64
14	70	65	64	60	45	40	68	73	75	62
15	51	53	66	60	52	62	60	54	63	65
16	50	38	49	44	50	55	68	61	60	60
17	53	45	61	45	57	50	70	77	60	55
18	56	53	63	63	68	40	74	62	40	60
19	64	66	64	60	67	64	55	55	60	66
20	60	48	55	60	67	57			62	64
21			51	45	56	54			57	63
22			58	50	58	53			71	$71\frac{1}{2}$
23			72	55	60	61			36	56
24			60	60	78	77			63	56
25			70	55	62	62			65	63
26			60	67	$72\frac{1}{2}$	65			56	57
27			62	62	57	60			66	62
28			60	60	43	54			36	57
29			80	60	70	49			51	75
30			60	53	70	57			45	48
31			60	60	85	65			50	54
32			60	60					55	45
33									60	65
34									71	$68\frac{1}{2}$
35									68	53
36									50	65
Mean . .	58.7	60.8	61.1	56.2	61.7	58.9	65.2	60.9	59.1	59.8

It is evident from an inspection of Table 6 that there is marked variation among the scorers in the evaluations of the same paper. The variation may be shown more strikingly by combining the five groups in two frequency tables, which show the distribution of scores for both the first and the second marking. This has been done in Table 7. The table shows that marks assigned during the first trial by 138 scorers ranged from 30 to 90, and that the marks assigned by the same scorers during the second trial ranged from 35 to 90. It is obvious, when such great differences occur in rating a pupil's paper, that little reliance can be attached to any given mark.

TABLE 7

DISTRIBUTION OF MARKS ASSIGNED A PUPIL'S ARITHMETIC PAPER
ON TWO DIFFERENT OCCASIONS BY THE SAME SCORERS

SCORE	FIRST TRIAL	SECOND TRIAL
85–89	3	1
80–84	3	1
75–79	4	4
70–74	17	7
65–69	17	21
60–64	43	43
55–59	22	22
50–54	18	23
45–49	3	12
40–44	5	3
35–39	2	1
30–34	1	
Totals . .	138	138
Mean . .	62.03	60.15

Summing up the evidence presented in Tables 6 and 7, we may conclude that in marking the type of examination paper in question, different raters do not agree with each other as

to what mark to assign, that they do not agree with their own first rating after an interval of four or five weeks, and that even the mean marks assigned by different groups do not agree. The cause of this disagreement may be sought in the lack of objectivity in marking the test used. For example, different raters vary in the extent to which they penalize the same error. One rater will mark a problem zero unless it is entirely correct both as to the answer and the process. Another rater will give full credit if the correct process is used in working the problem, even though the wrong answer is obtained. All sorts of standards for marking exist, ranging between these two extremes. Thus the mark assigned a given problem by a number of different raters might range from zero to perfect. The lack of objectivity in such a test also makes it difficult for a rater, after an interval of time has elapsed, to assign the same mark in his second rating. Even with the best intentions, he is likely to change the basis of marking. Indeed, the basis of marking may be changed unconsciously while one is marking a set of papers written for the same examination. This often accounts for the fact that two pupils, when they have given the same answer to a question, may be assigned different marks on these answers by the same rater.

Ashbaugh found in his experiment on this matter that the amount of variability in marking could be greatly reduced but not entirely eliminated by permitting the raters to work out a common or standard basis for marking. However, this method is not available to teachers separated from one another. Thus it is possible that a pupil would receive a mark of 30 on an arithmetic paper in one school, and a mark of 90 or more on the same paper if marked by a teacher in another school.

Standardized objective tests. Standardized objective tests have been developed largely in order to provide better

instruments for measuring the results of teaching than those afforded by the typical teacher's examination. Although these tests will be considered in detail in later chapters, we may here note, by way of contrast, how objective tests avoid some of the defects found in subjective tests. Our first illustration is taken from the Woody-McCall Mixed Fundamentals Arithmetic Test. The directions to the pupils for this test read as follows: "Get the right answer to as many examples as you can in twenty minutes. Do all work on the front or back of this sheet." The directions are followed by thirty-five exercises in the fundamentals, arranged in order of difficulty. The first six exercises, given below, illustrate the nature of the test.

(1)	(2)	(3)	(4)	(5)	(6)
Add			*Subtract*	*Multiply*	*Subtract*
22	2×3	$3)\overline{6}$	2	23	13
3			1	3	8

Our second illustration is taken from the Stanford Arithmetic Reasoning Test. The directions to the pupils read: "Find all the answers as quickly as you can. Write the answers on the dotted lines. Use the blank sheets to figure on." The directions are followed by forty exercises increasing in difficulty. Twenty minutes are allowed in which the pupils are to work as many of the exercises as they can. The nature of the exercises may be illustrated by the first five:

1. How many are 3 eggs and 2 eggs?.................. *Answer*____
2. Mary is 7 years old. How old will she be in 3 years? *Answer*____
3. A hen had 9 chicks, and 3 of them died. How many were left?...................... ....*Answer*____
4. Milk costs 8 cents a pint, and the milkman is going to raise the price 2 cents. What will it then cost?......*Answer*____
5. If you buy a pencil for 4 cents and pay for it with a dime, how much change should you get?............*Answer*____

Our third illustration is taken from the Stanford Sentence Meaning Test (Reading). The directions to the pupils are:

" Read each question and draw a line under the right answer."
The directions are followed by eighty exercises increasing in
difficulty. The first five will illustrate the nature of the
exercises :

1. Is milk white?.......... ....Yes   No
2. Do we sleep in beds?................................Yes   No
3. Is the day as dark as night?........................Yes   No
4. Is green a color?..... ....Yes   No
5. Is smoke always yellow?.......... .. ..Yes   No

It will be seen that the first two tests described above do
not differ in form from the ordinary subjective arithmetic
examinations. However, the objectivity of both tests is
insured by definite rules for scoring. A scoring key provides
the acceptable answers. Thus each exercise is scored either
as right or as wrong. In this way all scorers will obtain the
same results in scoring a given paper, provided they do not
make careless errors. The type of exercise described in our
third illustration differs from the traditional examination in
form and also in that it has definite directions for scoring.
Many other types of exercises used in objective tests are avail-
able, and they will be discussed in later chapters. Objective
standardized tests of different types have other important
features besides objectivity which make them more valuable
than subjective tests. These are discussed briefly in the fol-
lowing paragraphs.

Selection of content. The items which constitute such
tests as the foregoing are not selected arbitrarily by the
author of a test. On the contrary they are selected with the
greatest care. The general aim is to include in a test only
exercises that it is reasonable and desirable for pupils to be
familiar with. Catch questions and unimportant data have
no place in such a test. This usually means that the author
of a test must spend a great deal of time in examining courses
of study, textbooks, social usage, etc., in order to determine

what items to include in his test. The items must then be arranged in the best form possible as a preliminary form of the test. It must be given to hundreds of pupils in different grades to determine whether the items are of suitable difficulty, whether the language is unambiguous, etc. The results thus obtained are then carefully studied and the necessary changes and eliminations are made. It may be necessary to repeat this procedure several times before the content of the test is found satisfactory.

Scaling of items according to difficulty. In some tests the items selected are of equal difficulty. In this type of test a pupil's proficiency is determined by the number of items correctly done in a specified time. These are the so-called speed tests. In other tests the items are arranged in equal steps in order of their increasing difficulty. The tests on pages 23 and 24 illustrate the latter type. The difficulty of items in a test must be determined by actually giving the test to pupils in the grades for which it is intended. Thus if different items in a test have the same percentage of passes in a given grade, they may be said to be of equal difficulty for the pupils of that grade. Speed tests are of value in the drill subjects to determine the degree of skill or facility that a pupil has attained. Difficulty or power tests are of value in determining the range of a pupil's information in the subject concerned. Both types of tests are useful in diagnosing the weaknesses and strengths of a pupil, although special tests are also devised with diagnosis as their specific function.

Norms or standards. A third feature of standardized objective tests is that they provide norms or standards for comparison. Thus the Grade VI norm for the Woody-McCall Mixed Fundamentals Tests (page 23) is 22.5. Such norms are often provided for each age as well as for each grade. Norms are obtained by giving a test to thousands of pupils in different localities in order to determine the average

attainment of pupils in each of the grades or within the ages for which the test is devised. When the norms have once been established, the test may be used to determine the proficiency of a pupil or of a group of pupils in a way that is not ordinarily made possible by classroom examinations. Even though a teacher might succeed in making up a test in arithmetic that could be scored objectively, she still could not be sure whether her pupils, individually or as a class, had done as well as they should. She cannot decide whether her examination is easy or difficult except on the basis of her own opinion, which we have seen is not infallible because it is subjective. We often have an example of this subjectivity when there is a change of teachers. The teacher who leaves may have assigned uniformly low grades to the pupils; she may be followed by another teacher whose examinations indicate that the pupils have uniformly high grades. Consciously or unconsciously the grades assigned by the first teacher may have been affected by the hostility of pupils or parents, while those assigned by the second teacher may have been influenced by the desire to win their approval. On the other hand, the difference may be caused by the disagreement of these teachers as to the standards that should prevail. It is clear that norms provide an unbiased basis for determining the proficiency of a pupil or of a class.

Uniformity of administration. Most standardized tests specify that the directions to the pupils and the time allowed shall be uniform. Where this is the case, the examiner must adhere scrupulously to these requirements. It can readily be seen that a change in the wording of the directions to the pupils might make the test easier or harder than it was originally. It is clear also that this would affect the scores of the pupils so as to vitiate comparison with the norms. The scores are affected also when the examiner consciously or unconsciously varies the time limit. Even the addition or

subtraction of a fractional part of a minute may so change a pupil's score that it is useless to compare it with the norm. Such addition or subtraction would be like stopping a watch for a second while an athlete is attempting to set a record for the hundred-yard dash.

Arrangement of the content of a test. The items of an objective test may demand either recall or recognition. The most commonly used forms of each type are: [1]

    I. Recall
        a. Simple recall exercise
        b. Completion exercise
    II. Recognition
        a. Alternative response
        b. Multiple choice
        c. Matching
        d. Identification

The best arrangement of the content of a test depends upon the nature of the subject matter of the test, the amount of time to be given, the number of items to be included, etc. The method of scoring depends largely upon the type of test used. A common method is to base the score upon the number of items that are answered correctly. This method is sometimes modified when it is desired to weight the test in relation to other parts of the examination as a whole. Weighting may be accomplished by multiplying or dividing the number of items that are right by some predetermined number. Another means of scoring used particularly for the true-false type or for any other form of the alternative-response examination is the subtraction of the number of items answered incorrectly from the number answered correctly. The purpose of this procedure is to account for items answered correctly by chance.

[1] Concrete illustrations of each of these types as well as of other types or modifications of them will be found in later chapters.

Need for training in the use of standardized objective tests. Enough has been said concerning the nature of standardized objective tests to indicate the need for training in administering the tests before accuracy of measurement can be attained. No matter how good such a test is, the test alone does not insure accuracy of measurement. This depends upon the training and skill of the user. Objective measurement in education is, in this respect, similar to objective measurement in other fields. Thus the engineer, the electrician, or the chemist must learn how to use accurately the fine measuring instruments employed in his profession before his measurements will be accepted as reliable. Errors of measurement that are not due to inherent defects in the scale itself may be classified under two main types: (1) errors due to faulty observation and (2) errors due to faulty use of the measuring instrument. The first type is illustrated when a person who measures the width of a room miscounts the number of times he applies his yardstick. Similarly, in determining a pupil's score in an arithmetic test, the scorer may miscount the number of correct items. The second type of error is illustrated when mistakes are made in reading a gas or light meter through carelessness or lack of understanding. Errors of this type are frequent when inexperienced examiners attempt to use objective educational tests. As pointed out above, such errors occur when the examiner changes the directions or the time allowance. They occur also when the examiner fails to understand the method of computing the score for the test.

Inaccuracy in scoring standardized tests. When tests are scored by teachers with little or no training, inaccuracies are more common than is generally realized. In order to check on the type and frequency of errors by untrained teachers, the writer conducted an experiment with 47 normal-school seniors in a class in educational tests and measurements. Most of these seniors had previously had some teaching

TABLE 8

SHOWING ERROR IN SUBJECT AGES IN EACH OF NINE SUB-TESTS IN THE STANFORD ACHIEVEMENT TEST, ADVANCED, DUE TO ERRORS IN SCORING

SCORER	TEST								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	11–10 12–7				10–10 12–9				
2		11–11 14–8	11–6 13–2						
3		10–6 13–0				13–0 12–2			
4						17–8 15–11	15–9 14–4	17–1 15–4	
5		8–9 9–11				8–9 9–8	8–9 9–8	8–9 11–0	
6		11–11 13–2				10–4 11–11	12–6 11–2		10–11 10–10
7		17–1 13–5		8–11 11–6				16–8 14–0	
8								15–2 11–0	
9						10–11 12–5	10–5 12–2	15–2 19–11	
10			16–3 10–4						
11		9–7 12–11	12–1 13–3	12–8 17–7				13–5 11–7	
12				11–8 10–9					
13								8–6 8–10	
14								8–6 10–0	
15								9–0 10–11	

NOTE. In the above table the upper score is in each case in error and the lower score is correct.

experience as well as some experience in the use of standardized tests. As part of their training they were required to score two or more Stanford Achievement Tests. Before they began, the method of scoring was explained and discussed until every student was satisfied that he understood the procedure. Each member of the class then scored a booklet containing the responses of a pupil who had taken the test. The booklets were collected and later were rescored by the writer. Of the 47 booklets thus scored, 15 contained errors in scoring in one or more of the nine sub-tests. Table 8 shows for each of the nine sub-tests the discrepancies between the erroneous and the correct scores, expressed as subject ages.

From the preceding table it will be seen that 15 of the 47 scorers made 33 errors in scoring, which resulted in errors in subject ages. The 33 subject ages range in error from one month to five years and eleven months. With three exceptions the errors range in size from eleven months to five years and eleven months. The following tabulation shows the method of scoring each of the nine sub-tests, and the type and frequency of error :

Test 1. (Number right times two)
Error in counting the number right : 1
Test 2. (Right minus wrong)
Omitted items counted as wrong : 3
Omitted items counted as right : 1
Only omitted items counted as right : 1
Test 3. (Number right)
Error in counting the number right : 5
Omitted items included with number right : 1
Test 4. (Number right times four)
Error in arithmetic, such as 20 times 4 equals 100 : 1
Error in counting number right : 1

Test 5. (Number right times four)
     Error in counting number right: 1
Test 6. (Right minus half of number wrong)
     Omitted items counted as wrong: 5
Test 7. (Right minus half of number wrong)
     Omitted items counted as wrong: 2
     Omitted items counted as right: 1
Test 8. (Right minus wrong)
     Error in arithmetic, such as 60 minus 20 equals
       30: 3
     Number wrong not subtracted: 1
Test 9. (Number right times four)
     Error in counting number right: 1

In the tabulation above, the number of errors is given after the type of error. Of the 28 errors, 15 may be ascribed to failure to understand the directions for scoring; 9 may be classified as observational errors in counting; and 4 as errors in arithmetic. In addition to the 28 errors listed above, there were 5 errors in converting the point scores into subject ages by means of the table in the manual of directions, such as reading the subject ages from the wrong column.

These 15 scorers, after their errors in scoring had been ascertained, were called for individual conference, and their errors were pointed out to them. The method of scoring was gone over once more for those tests where errors in scoring were found. The whole class of 47 were then given similar booklets to score, after they had been cautioned about the difficulties in scoring listed above. The booklets were checked by the writer in the same manner as the first set. This time material errors in scoring, of about the same type and size as those listed in Table 8, were made only by Scorers 2, 3, 8, 11, and 14. In other words, additional training resulted in cutting down the errors very materially, though it

did not entirely eliminate them. Doubtless further training would reduce the errors even more. The writer is led to believe that the tabulations presented above are fair indices of the degree of accuracy that we may expect from untrained scorers and of the amount of training required to increase it. He does not base his conclusion entirely on his own experience but also on two similar studies of the degree of accuracy in scoring among teachers. One of these studies was made by Franzen and Hanlon,[1] and the other by Pintner.[2] They show errors in scoring of about the same type and extent as those just described.

Instead of becoming discouraged by the types and frequency of the errors indicated in these studies, beginners should regard them as a warning and a challenge. The novice in teaching probably errs no more in the application of standardized tests than he does in the application of other tools of the teaching profession. Indeed, in the use of standardized tests the possibility of discovering and stating errors in quantitative terms is a distinct advantage. We have fewer misgivings about the use of other complex teaching devices, such as supervised study, the project method, and methods of teaching silent reading, in part perhaps because it is harder to get precise measurements of mistakes made by teachers who use them.

General and specific training. To become expert in the use of a standardized test, the novice should have sufficient knowledge of the theoretical and technical principles of the test to appreciate its nature and function. In addition to this, he should have specific training in the actual administration of the test or tests which he expects to use. Thus

[1] Raymond H. Franzen and W. H. Hanlon, *The Program of Measurement in Contra Costa County.* Standard Print, Martinez, California; 1923.

[2] Rudolf Pintner, "Accuracy in Scoring Group Intelligence Tests," in *The Journal of Educational Psychology,* Vol. XVII, pages 470–475; October, 1926.

Terman [1] has shown that a few weeks' training will enable a teacher to use the Binet Tests with a fair degree of accuracy. Dickson and Martens [2] have similarly shown how this can be accomplished in large groups of teachers who are learning how to use these tests. Similar care should be exercised in training examiners to use other tests.

EXERCISES

1. List other examples, not mentioned in the text, of objective measurement in the physical world.

2. Let each member of the class mark the same arithmetic paper written by some pupil. Compare the variability with the examples given in the text.

3. Obtain a set of examination questions written by a teacher. Try to determine the relative difficulty of the various questions. After the examination has been given to the pupils and the papers have been scored, compare your estimates with the actual difficulty of the questions as it is indicated by the number of pupils answering each question satisfactorily.

4. Find examples of teachers' examinations where the questions are ambiguous; others where the information asked for is not important.

5. If the pupils seem not to understand the directions for a standardized test, should the examiner attempt to explain or simplify them?

6. Obtain samples of the standardized tests discussed in this chapter. Attempt to classify them as to type, according to the scheme given on page 27.

7. Find from daily life examples of errors of measurement due to faulty observation; errors due to faulty use of the measuring instrument.

8. Divide the class into groups, each containing five or six students. Let each group score a standardized test paper without seeing the scores obtained by other members of the group. If dif-

[1] L. M. Terman, *The Measurement of Intelligence*, pages 107–109. Houghton Mifflin Company, Boston; 1916.

[2] V. E. Dickson and E. H. Martens, "Training Teachers for Mental Testing in Oakland, California," in *Journal of Educational Research*, Vol. VII, pages 100–108; February, 1923.

ferences in scoring are found, note the causes, such as errors in counting items, errors in arithmetical computations, misunderstanding of directions for scoring, etc.

References

BUCKINGHAM, B. R. *Research for Teachers,* Chapter VI. Silver, Burdett & Co., New York; 1926.

GILLILAND, A. R., and JORDAN, R. H. *Educational Measurements and the Classroom Teacher,* Chapter IV. Century Company, New York; 1924.

HULTEN, C. F. "The Personal Element in Teachers' Marks." *Journal of Educational Research,* Vol. XII (June, 1925), pages 49–55.

KELLY, F. J. *Teachers' Marks, Their Variability and Standardization* (Contributions to Education, No. 66). Teachers College, Columbia University, New York; 1914.

McCALL, WILLIAM A. *How to Measure in Education,* Chapter I. The Macmillan Company, New York; 1922.

MONROE, W. S. *The Theory of Educational Measurements,* Chapter II. Houghton Mifflin Company, Boston; 1923.

——, DeVOSS, J. C., and KELLY, F. J. *Educational Tests and Measurements,* Chapter I. Houghton Mifflin Company, Boston; 1924.

RUCH, G. M. *The Objective or New-Type Examination,* Chapter III. Scott, Foresman & Co., Chicago; 1929.

STARCH, DANIEL. *Educational Psychology* (Revised), Chapter XXIII. The Macmillan Company, New York; 1927.

THORNDIKE, E. L. "Measurement in Education." *Teachers College Record,* Vol. XXII (November, 1921), pages 371–379.

TRABUE, M. R. *Measuring Results in Education,* Chapters I and II. American Book Company, New York; 1924.

CHAPTER THREE

STATISTICAL METHODS: TABULATION AND CLASSIFICATION

Necessity for tabulation and classification.[1] In order that the mind may more easily grasp and retain the results of measurement or may remember collections of numerical data of varying magnitudes, it is usually necessary to tabulate or classify the data in some systematic way. That is, when more than a few measures are involved, the mind is unable to retain all the individual measures or facts so as to observe the significant trend, or to make comparisons with a similar group or groups. An illustration of this may be seen in the election of a president of a club. Suppose Smith, Jones, and Brown are candidates and have been voted on by the members of the club. A committee is usually appointed to collect the votes and determine who has the majority. This may be done by writing the names of the three candidates on a sheet of paper and checking, or tallying, the votes for each candidate as follows:

SMITH   ~~丗~~ ~~丗~~ ~~丗~~ ~~丗~~ ~~丗~~ ~~丗~~ ~~丗~~ /   (36)
JONES   ~~丗~~ ~~丗~~ ~~丗~~ ~~丗~~ ~~丗~~ ~~丗~~ ///   (33)
BROWN ~~丗~~ ~~丗~~ ~~丗~~ ~~丗~~ ~~丗~~ ~~丗~~ ~~丗~~ ~~丗~~ ~~丗~~ (45)

It will be noted in this tabulation the tallying is done in groups of five, the first four tally marks being written vertically and the fifth placed diagonally to tie them together. This customary way of checking or tallying facilitates counting. Thus we see that the three candidates have 36, 33, and 45 votes each. A similar procedure is used in tabulating educational measurements, though there are several other

[1] In connection with the study of this chapter there should be an abundance of opportunity for the tabulation and classification of statistical data. Actual test scores would be most valuable for this purpose. Similar provision for practice in computing the measures discussed in Chapter IV should be made.

procedures available which we shall consider first. Suppose we have given a reading test to a group of pupils. We may then tabulate the results alphabetically as follows:

TABLE 9

SHOWING POSSIBLE RESULTS OF A READING TEST, ARRANGED
ALPHABETICALLY BY NAMES

NAME	SCORE	NAME	SCORE
1. Allen	80	17. Kitts	70
2. Ashley	100	18. Lawrence	74
3. Bailey	96	19. Lewis	95
4. Black	114	20. Martin	70
5. Blewett	105	21. McCarty	106
6. Brown	93	22. Merry	91
7. Conlon	74	23. Myers	100
8. Connell	56	24. Perry	85
9. Davis	88	25. Platt	90
10. Dawes	108	26. Richards	100
11. Dickson	52	27. Ruger	88
12. Donovan	52	28. Steele	45
13. Doty	95	29. Stickle	103
14. Fisher	94	30. Stork	82
15. Horne	77	31. Warren	70
16. Johnson	104	32. Wayne	69

A tabulation such as the above is convenient for finding quickly the score of a given pupil. The same result may, of course, be obtained by using an alphabetical card index containing a card for each pupil. For quick reference some such system is necessary. The method of record-keeping illustrated above does not, however, make it easy to determine the relative ranks of the pupils concerned. If one is rather familiar with the meaning of scores, this is not so necessary. However, in order to show where a given pupil ranks in his group, the names in the above tabulation may be rearranged in the order of magnitude of score, as follows:

TABLE 10

SHOWING POSSIBLE RESULTS OF A READING TEST, ARRANGED
IN ORDER OF SCORES

Name	Score	Name	Score
1. Black	114	17. Davis	88
2. Dawes	108	18. Ruger	88
3. McCarty	106	19. Perry	85
4. Blewett	105	20. Stork	82
5. Johnson	104	21. Allen	80
6. Stickle	103	22. Horne	77
7. Ashley	100	23. Conlon	74
8. Myers	100	24. Lawrence	74
9. Richards	100	25. Kitts	70
10. Bailey	96	26. Martin	70
11. Doty	95	27. Warren	70
12. Lewis	95	28. Wayne	69
13. Fisher	94	29. Connell	56
14. Brown	93	30. Dickson	52
15. Merry	91	31. Donovan	52
16. Platt	90	32. Steele	45

When the scores have been arranged in this order, one can obtain a very definite notion of a pupil's standing in relation to the group in which he is enrolled. One can also determine very easily how many pupils have scores above or below a given point. For example, the table above indicates that sixteen pupils, or one half of the class, have scores as high as 90 or more, while the other half have scores as low as 88 or less. The average of these scores (90 and 88) is 89 and is called the midscore.[1] If there had been an uneven number of scores, such as thirty-three, the seventeenth score, counting in from either side, would be the midscore. The midscore is average, though not identical with the arithmetical

[1] The midscore is sometimes confused with the median. The median, however, is calculated from the frequency distribution, as will be shown in Chapter IV.

average, and may be used similarly in comparing one group
with another.   In the same manner, the best fourth of the
pupils may be found to have scores of 100 or more, while
the poorest fourth have scores of 70 or less.   This procedure
aids in giving meaning to scores, especially if the user is not
familiar with the test.

The frequency table.   However, when large numbers of
pupils are examined, the above method of ranking the scores
in order of magnitude becomes rather slow and tedious.
Suppose, for example, pupils receive the following scores in a
reading test : 94, 86, 101, 68, 87, 83, 93, 71, 93, 96, 81, 115,
72, 111, 98, 120, 75, 97, 72, 85, 103, 85, 82, 76, 101, 85, 99,
116, 125, 87, 97, 90, 98, 70, 94, 106, 121, 66, 69, 100, 80, 96,
93, 114, 105, 74, 56, 113, 88, 108, 77, 64, 95, 88, 94, 85, 81,
94, 109, 79, 106, 77, 75, 104, 101, 80, 70, 81, 97, 84, 95, 70,
76, 96, 91, 78, 100, 90, 75, 85, 105, 90, 84, 100, 88, 74, 63,
92, 103, 65, 96, 61, 60, 69, 101, 115, 85, 100, 95, 70, 105, 72,
105, 80, 94, 97, 71.   It is obvious that considerable time would
be consumed in arranging the scores in order of their magni-
tude as in the table illustrated above.   We may instead pro-
ceed to tabulate the scores in the form of a frequency table,
as illustrated below.   The first score given above, 94, will
obviously fall in the interval 90–94 ; the second score, 86,
will fall similarly in the interval 85–89.   Continuing in this
way we obtain a frequency table such as the one shown on the
following page.

A frequency table such as Table 11 makes it possible for
us to observe the general trend of the distribution of
scores in a manner that is not possible with a large number
of scores that are not so tabulated.   It shows at a glance the
number of pupils whose scores fall within a given group.
Thus one pupil obtains a score between 125 and 129 ; the
next highest are two pupils with scores between 120 and
124 ; next are three pupils with scores ranging between 115

TABLE 11

FREQUENCY TABLE, SHOWING THE DISTRIBUTION OF SCORES IN
GRADE VIII ON THE HAGGERTY READING EXAMINATION, SIGMA 3

SCORE	TABULATION	FREQUENCY
125–129	/	1
120–124	//	2
115–119	///	3
110–114	///	3
105–109	⁄⁄⁄⁄ ///	8
100–104	⁄⁄⁄⁄ ⁄⁄⁄⁄ /	11
95–99	⁄⁄⁄⁄ ⁄⁄⁄⁄ ////	14
90–94	⁄⁄⁄⁄ ⁄⁄⁄⁄ ///	13
85–89	⁄⁄⁄⁄ ⁄⁄⁄⁄ //	12
80–84	⁄⁄⁄⁄ ⁄⁄⁄⁄	10
75–79	⁄⁄⁄⁄ ////	9
70–74	⁄⁄⁄⁄ ⁄⁄⁄⁄ /	11
65–69	⁄⁄⁄⁄	5
60–64	////	4
55–59	/	1
Total . .		107

and 119; etc. In a word, such a table shows the frequencies
with which scores occur in the different groups.

It may be well to point out at this time that a trait such as
the one tabulated in Table 11 is technically known as a vari-
able. Any trait that may appear in differing amounts may
be defined as a variable. Thus height, weight, age, and
achievement test score are examples of variables. Variables
may be either continuous or discrete (discontinuous). In
a continuous variable measurements may be made without
the occurrence of breaks or gaps; for example, such vari-
ables include height, weight, and scores in achievement and
intelligence tests. Thus, reading achievement would be
classified as a continuous variable. It is true that the read-
ing scores tabulated in Table 11 are stated in terms of whole

numbers, and thus appear to leave gaps. However, with a sufficiently fine measuring instrument it would be possible theoretically to state the differences in reading ability in fractions of the unit of measurement, such as 74.4, 74.9, etc. What has been said of reading appears to be true of other school subjects when pupils are tested with adequate measuring instruments. Indeed it seems true of all human traits that are measurable.

The second type, or discontinuous variable, is one in which gaps may occur when measurements are made. This type would be illustrated by a tabulation giving the number of pupils taught by different teachers. In such a distribution breaks or gaps would occur. For example, a teacher could not have 29¾ pupils in a class. While discontinuous variables are common enough in general statistics, it is clear that in measuring the achievement and ability of school children, the variables will be continuous.

The class interval of a distribution. The class interval of a frequency table may be defined as one of the equal parts into which the table is divided for purposes of tabulation. In Table 11 the class interval is stated as 50–54, 55–59, etc., and thus appears to range over four units. However, the range is really five units. That is, the interval 50–54 ranges from 50 up to, but not including, 55. Thus it might be written 50–54.9999 . . . 9. This is true also of the other intervals. To avoid ambiguity and confusion in tabulating measures or scores, the intervals are not stated as 50–55, 55–60, 60–65, etc. Suppose, for example, a score of 55 was obtained. Such a score could be placed equally well in either the first or the second class interval.

The size of the class interval may be determined arbitrarily, taking into account both convenience in tabulating and the interval that will most truly represent the original data. Ordinarily it is not advisable to have less than twelve classes

or more than eighteen or twenty. Fewer than twelve classes may obscure the magnitude of the individual differences represented; more than eighteen or twenty may result in a table that is too unwieldy to handle. The number of class intervals may be estimated by finding the difference between the lowest and the highest measures or scores, and by dividing this difference by a number which will yield the desired number of classes. For example, an inspection of the unclassified reading scores on page 38 shows that they range from 56 to 125, or that there is a difference of 69 between the lowest and the highest. Dividing this number by 5 yields a quotient of 13.8; so an interval of five will result in 15 classes, because each of the two extreme scores will require a whole class interval. For statistical purposes it is assumed that the midpoint of a class interval best represents the magnitude of the measures grouped opposite that interval. Thus the midpoint for class interval 100–104 in Table 11 would be 102.5.

Frequency surfaces. Often it is desirable to present the facts of a frequency table graphically. This may be done by plotting the table in the form of a frequency surface. Two methods for plotting this surface are in common use. The facts may be represented in the form of a histogram (or column diagram), or in the form of a frequency polygon. These methods are illustrated in Figures 5 and 6, which represent graphically the facts in Table 11.

The following is the procedure in constructing the first type, or histogram (see Figure 5 on page 42) :

1. Draw a base line. Mark off the class intervals on this base line.
2. Draw a vertical line at the left end of the base line. On this vertical line indicate the number of cases at equal intervals.

3. Place a dot above the midpoint of each interval and opposite the number of cases indicated for that interval.
4. Draw a horizontal line through each dot, joining the extremes of each interval.
5. Draw lines to the base from the extremes of these horizontal lines, thus forming a series of columns.

In constructing the frequency polygon, use a similar procedure (see Figure 6 on page 44) :

1, 2, 3.   Follow steps 1, 2, and 3 used for the construction of the histogram shown in Figure 5.
4. By straight lines connect all the dots which you indicated in step 3.

Fundamentally, the histogram and the frequency polygon represent the same facts concerning a frequency distribution. The histogram represents the frequencies of each interval
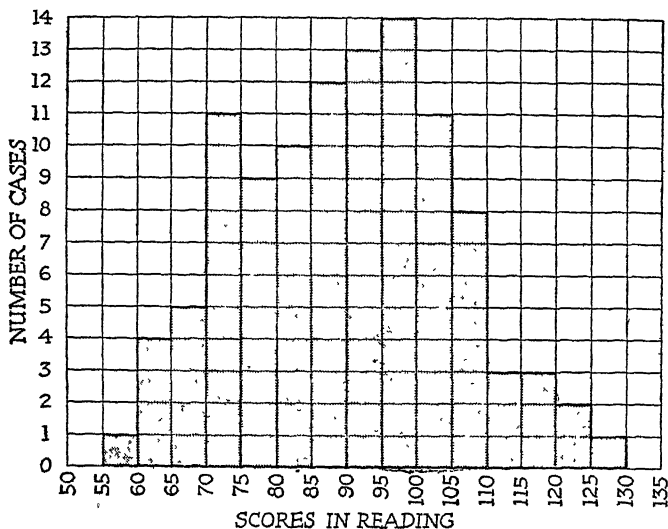
FIG. 5.   Histogram, or column diagram.   This figure represents the same data as does Table 11.

by a series of rectangles or columns, the height of each being determined by the number of cases. This makes it appear that there are gaps or breaks between the frequencies of two adjacent intervals, which we have seen is not true of continuous variables. This difficulty is avoided in the frequency polygon, which more accurately represents the continuous nature of such a variable as the one tabulated in Table 11.

The histogram and the frequency polygon are sometimes called frequency curves. Some authorities reserve this term, however, for the "smoothed" frequency polygon. The frequency curve is drawn in the same way as the frequency polygon but is made more symmetrical. This is done by smoothing the irregularities which appear because of too small a sampling or too large an interval, or because of errors in measurement. The form of such a smoothed curve approaches the theoretical probability curve discussed later in this chapter.

The normal frequency curve. This curve is also known as the normal surface of frequency, the normal probability curve, and the Gaussian curve. It is the generalized theoretical curve, of which histograms and frequency polygons are concrete examples. In form it is a symmetrical, bell-shaped curve such as the one illustrated by Figure 8. It has been found that when most human traits are measured accurately and in large numbers the measures tend to approximate the form of distribution that is represented by the normal frequency curve. This curve may also be obtained by tossing coins and then tabulating the number of times that heads and tails appear. The theoretical normal curve may be constructed with comparative ease by expanding a binomial, such as $(1 + 1)^{10}$. This method would result in the values that are tabulated in Table 12 under the heading " Theoretical Distribution." The sum of these values is 1024. If we take ten coins — for example, ten pennies —
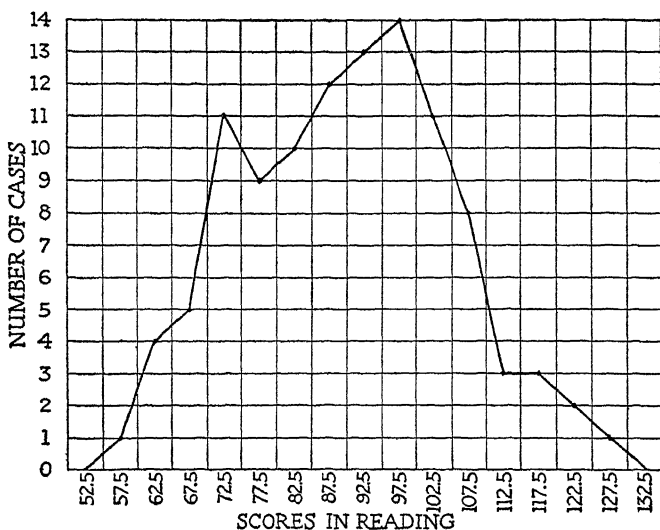
Fig. 6. Frequency polygon. This figure represents the same data as do Table 11 and Figure 5.

and toss them 1024 times, counting and tabulating the number of heads and tails that appear each time, we shall obtain a distribution similar to the theoretical distribution. The figures listed under the heading " Actual Distribution " were obtained in this way by the writer.

Measurements of human traits, when tabulated or plotted in the form of a frequency polygon, tend, under appropriate conditions, to be distributed according to the normal frequency curve. It is therefore obvious that this curve is a useful means of checking measurements of any given trait. If the distribution of such measurements shows decided deviation from the normal curve, we should carefully investigate the causes of the variation.

Skewed distributions. Frequency distributions may sometimes show a tendency to lean to one side. This condition

TABLE 12

THEORETICAL AND ACTUAL DISTRIBUTION OF HEADS AND TAILS
OBTAINED BY TOSSING TEN PENNIES 1024 TIMES

H	T	THEORETICAL DISTRIBUTION	ACTUAL DISTRIBUTION
0	10	1	1
1	9	10	9
2	8	45	44
3	7	120	122
4	6	210	211
5	5	252	263
6	4	210	207
7	3	120	123
8	2	45	37
9	1	10	6
10	0	1	1
Totals . . .		1024	1024
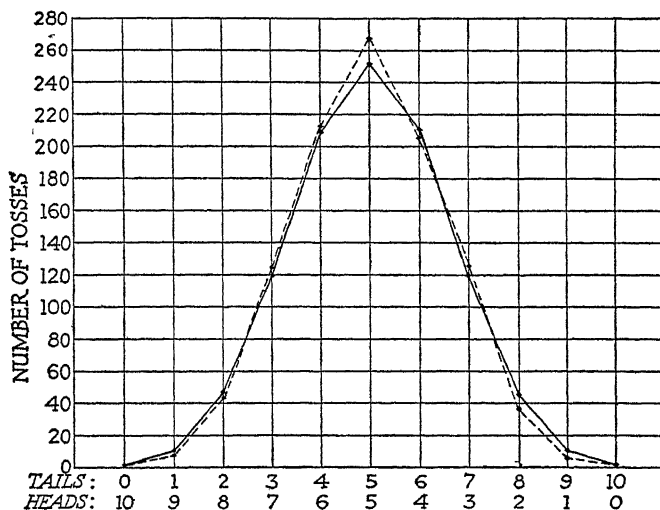
FIG. 7. Curve showing theoretical and actual distribution of heads
and tails in 1024 throws of 10 pennies. Data from Table 12.
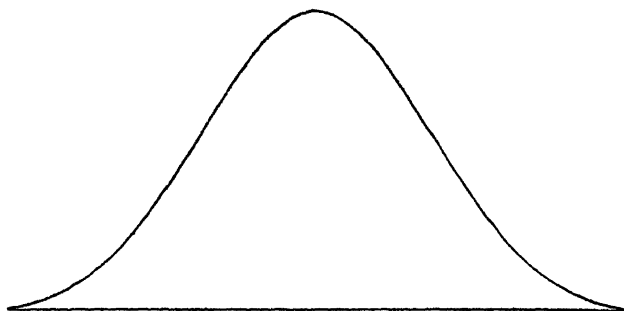
Fig. 8. A normal probability curve which illustrates the symmetrical bell-shaped curve obtained when an infinitely large number of measurements are obtained under appropriate conditions.

has been termed *skewness*, and it is present when the scores are massed between the middle of the distribution and the upper or lower end. This type is illustrated in Table 13 by two distributions of school marks, one for English and the other for mathematics.

TABLE 13

DISTRIBUTION OF SCHOOL MARKS IN ENGLISH AND MATHEMATICS

LETTER GRADE	ENGLISH	MATHEMATICS
A	3	31
B	12	36
C	16	18
D	36	10
F	33	5
Totals . . .	100	100

Table 13 shows that in English 3 students in 100 received a mark of A, 12 received B, 16 received C, etc. In mathematics, 31 students in the same group received A, 36 received B, 18 received C, etc. From these marks we may conclude that the English instructor used a much more severe standard than the mathematics instructor. Such differences in

standards occur frequently among teachers and, of course, make it impossible to compare marks earned in one course with those earned in another. When these tables are plotted in the form of histograms, they appear as indicated in Figures 9 and 10.
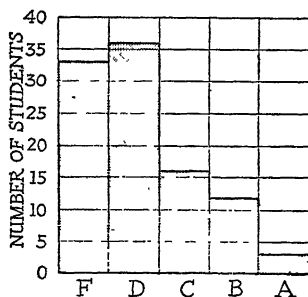
FIG. 9. Negative skewness.   FIG. 10. Positive skewness.
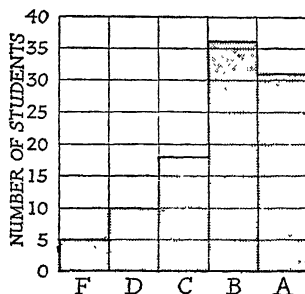
Various causes operate to produce skewness in distributions. As has been suggested above, errors in measurement (when the errors are in one direction) result in skewness. The improper selection of individuals for measurement is another cause. For example, when the group that is measured is small, chance may produce either positive or negative skewness. Again it may happen that the individuals selected are not characteristic of the whole group which they represent.

Bi-modality and multi-modality. Other distributions that vary from the normal type may be listed. We shall discuss only one of these, because the others do not come within the scope of this book. The normal frequency distribution may be called uni-modal; that is, having one mode. The mode is an average which is identical with both the median and the average in a perfectly normal frequency distribution. It may be found by rather elaborate computations or by inspection. When it is found by inspection, it is called the crude mode. In this case the mode may be determined by noting

TABLE 14

DISTRIBUTION OF READING SCORES IN SENIOR AND FRESHMAN HIGH SCHOOL
    CLASSES RESULTING IN BI-MODALITY WHEN THE TWO DISTRIBUTIONS
    ARE COMBINED

SCORE	SENIOR	FRESHMAN	COMBINED
130–134	1		1
125–129	2		2
120–124	5		5
115–119	6	2	8
110–114	7	5	12
105–109	9	10	19
100–104	17	12	29
95–99	21	16	37
90–94	15	16	31
85–89	15	23,	38
80–84	6	14	20
75–79	3	12	15
70–74	3	8	11
65–69		5	5
60–64		2	2
55–59		2	2
50–54		1	1

the interval in the frequency distribution that has the larg-
est number of cases. This interval contains the mode. If
accuracy is not demanded, the mode is a useful inspectional
average to indicate the central tendency.

It may happen that two intervals, each containing a large
number of cases, are separated by one or more intervals con-
taining fewer cases. In this event the distribution is said
to be bi-modal. This distribution is frequently caused by
the combination, accidental or otherwise, of two distinct
groups, as illustrated in Table 14. This table shows the
distribution of scores obtained on the Haggerty Reading
Examination, Sigma 3, by the senior and freshman classes in
a high school. When these two distributions are combined,

two modes appear, one in the class interval 95–99, and the other in the class interval 85–89.

Bi-modality is also called multi-modality, but this latter term may also mean that a given distribution includes more than two modes. Sometimes the combination of distinct groups, instead of producing two or more modes, may result as above in wider separation of the extremes and may also flatten the peak of the frequency polygon. When multi-modality is indicated in a distribution, it is desirable to examine the data and to determine, if possible, the cause for the deviation from normality.

EXERCISES

1. Obtain the test scores in some subject for a class of from twenty to thirty pupils. Classify the scores by arranging the names of the pupils in alphabetical order and writing each pupil's score after his name.

2. Arrange the same scores in order of magnitude, starting with the highest, as shown in the example on page 37.

3. Make two frequency tables from the scores given on page 38, using a class interval of 3 for the first, and of 10 for the second. Compare these tables with Table 11.

4. Give examples of variables other than the ones given in the text. Classify them as continuous or discontinuous variables.

5. From the data obtained for Exercise 1, plot a histogram and a frequency polygon.

6. Construct the normal curve by expanding the binomial $(1 + 1)^6$. Compare this curve with the curve obtained by tossing 6 pennies 64 times and tabulating the frequencies with which heads and tails appear.

7. Using the material in Table 14, plot frequency polygons for the seniors, for the freshmen, and for the combined groups.

References

BUCKINGHAM, B. R. *Research for Teachers*, Chapter II. Silver, Burdett & Co., New York; 1926.

GARRETT, HENRY L. *Statistics in Psychology and Education*, Chapters I and II. Longmans, Green & Co., New York; 1926.

GREGORY, C. A.   *Fundamentals of Educational Measurement,* Chapter IX.
D. Appleton & Co., New York; 1922.

LINCOLN, E. A.   *Beginnings in Educational Measurement,* Chapter III.
J. B. Lippincott Company, Philadelphia; 1924.

McCALL, WILLIAM A.   *How to Measure in Education,* Chapters XII and
XIII.   The Macmillan Company, New York; 1922.

ODELL, C. W.   *Educational Statistics,* Chapters I and II.   Century Com-
pany, New York; 1925.

OTIS, ARTHUR S.   *Statistical Method in Educational Measurement,* Chapters
III and IV.   World Book Company, Yonkers-on-Hudson, New York;
1925.

RUGG, H. O.   *A Primer of Graphics and Statistics for Teachers,* Chapter II.
Houghton Mifflin Company, Boston; 1925.

THURSTONE, L. L.   *The Fundamentals of Statistics,* Chapters I, II, and III.
The Macmillan Company, New York; 1925.

TRABUE, M. R.   *Measuring Results in Education,* Chapters V and X.
American Book Company, New York; 1924.

YULE, G. UDNY.   *An Introduction to the Theory of Statistics,* Chapter VI.
C. Griffin & Co., London; 1922.

CHAPTER FOUR

Statistical Methods: Central Tendency, Percentiles, Variability, Correlation, Reliability

I. MEASURES OF CENTRAL TENDENCY

The need for measures of central tendency. In working with statistical data it frequently becomes desirable to compare one group of measurements with another. We may desire to compare men and women as to height, boys and girls as to achievement in some school subject, men and women teachers as to salary, etc. We cannot do this on the basis of individual comparisons because the two individuals being compared may not be typical of the group they represent. Thus, if we say that Mr. A is taller than Miss B, someone may recall that Miss C is taller than Mr. A. Such a matching process might be continued endlessly without determining which sex is the taller. If, however, we obtain measurements of a great many men and women and compute the average height of each sex group, we are able to make very definite comparisons. The average may be used as a summary statement for the group measured.

We have already seen that in many important measurable traits human beings differ from each other. It therefore frequently becomes desirable to have an average measure for a specified trait so that we may know how any individual compares with his group. Teachers constantly make such comparisons when they attempt to determine whether a pupil is above or below the class average in a given subject. Everyone is familiar with the use of the average, or *arithmetic mean*, as it is technically called, for purposes such as those listed above. In addition to the mean there are several other measures for determining the central tendency of a group. Of these, we need concern ourselves with only

51

the *median* and the *mode*, which, together with the mean are the measures of central tendency most commonly used in educational statistics.   In the preceding chapters reference has been made to the three measures of central tendency — the mean, the median, and the mode.   However, we need to consider in more detail the uses of these measures and the methods for computing them.

The mean.   Everyone is familiar with the common method of computing the mean : To find the average age of a group of 30 pupils, we may write their ages in a column, and divide the sum of the column by the number of pupils (30) ; the quotient is the mean, or average.   This procedure is satisfactory in dealing with a small group.   Suppose that instead of 30 pupils we have several hundreds or even thousands of pupils.   Obviously the method of finding the mean just described would become very laborious and tedious.   For this reason it is sometimes more desirable to use certain time-saving methods in computing the mean.

Computing the mean from a frequency distribution.   Several methods are available for computing the mean from a frequency distribution.   The first method may be illustrated by data given in Table 15, which shows the distribution of heights among 202 women.

In Table 15 each height has been multiplied by the number of individuals of that height.   Beginning at the top of the table, we see that 1 person is 69 inches tall, and that the height, when multiplied by the frequency, yields a product of 69.   In the second item, 3 persons are each 68 inches, the height multiplied by the frequency yielding a product of 204 inches.   The next 8 persons are each 67 inches, and the total is 536 inches; etc.   Adding the last column, we obtain the sum of 12,758, in which total is included the 202 heights tabulated.   Now, to find the mean or average height from this total, we merely divide 12,758 by 202, obtaining the

TABLE 15

ILLUSTRATING THE METHOD OF COMPUTING THE MEAN FROM A FREQUENCY
DISTRIBUTION BY MULTIPLYING EACH MEASURE BY ITS FREQUENCY

X (Height in Inches)	f (Frequency)	fX
69	1	69
68	3	204
67	8	536
66	19	1254
65	26	1690
64	32	2048
63	36	2268
62	29	1798
61	24	1464
60	15	900
59	6	354
58	2	116
57	1	57
Totals . . .	202	12758

quotient 63.15. The formula for computing the mean by
this method may be stated as: $M = \dfrac{\Sigma fX}{N}$.

In this formula:

M is the abbreviation for " mean."

The Greek capital letter Σ (sigma) is used to mean "sum
of."

f is the abbreviation for " frequency."

X designates the variable, which in the example above
is "height."

N signifies the number of cases.

The formula may be applied by substituting the values
obtained in the example on page 52. By this process we
obtain: $M = \dfrac{12758}{202} = 63.15$.

In the above procedure we have assumed that the measurements of height are in terms of whole inches, or that the variable is discontinuous.   If we consider the variable as continuous, a slight change in the procedure must be made. Before multiplying a given height — for example, 66 — by its given frequency 19, we would change the multiplicand to 66.5.   In this method it is recognized that the 19 individuals mentioned range in height from 66 up to but not including 67 inches, and that as a result the midpoint 66.5 becomes the average for the group.

When statistical data are tabulated in a frequency table with an interval of any given size, we may proceed as illus-

TABLE 16

ILLUSTRATING THE METHOD OF COMPUTING THE MEAN FROM A FREQUENCY TABLE BY MULTIPLYING THE MIDPOINTS OF EACH INTERVAL BY THE FREQUENCIES

CLASS INTERVAL	X (Midpoint of Interval)	f	fX
125–129	127.5	1	127.5
120–124	122.5	2	245.0
115–119	117.5	3	352.5
110–114	112.5	3	337.5
105–109	107.5	8	860.0
100–104	102.5	11	1127.5
95–99	97.5	14	1365.0
90–94	92.5	13	1202.5
85–89	87.5	12	1050.0
80–84	82.5	10	825.0
75–79	77.5	9	697.5
70–74	72.5	11	797.5
65–69	67.5	5	337.5
60–64	62.5	4	250.0
55–59	57.5	1	57.5
Total (N) . . .		107	9632.5

trated in Table 16.[1]   This method is essentially the same as
the one described for Table 15.   When intervals are given,
the midpoint of each interval is multiplied by its frequency.
The remaining steps also are the same as those followed in
Table 15.

Applying the formula $M = \dfrac{\Sigma f X}{N}$, we find that the mean

for the above distribution is 90.02.   It is clear that the pro-
cedure, aside from the difference in the interval, is the same
as that used in our first example (page 53).   The advantages
of the second method are limited, because when it is applied

TABLE 17

ILLUSTRATING THE SHORT METHOD FOR COMPUTING THE
ARITHMETIC MEAN

CLASS INTERVAL	f	d	fd
125–129	1	7	7
120–124	2	6	12
115–119	3	5	15
110–114	3	4	12
105–109	8	3	24
100–104	11	2	22
95–99	14	1	14
90–94	13	0	106
85–89	12	−1	−12
80–84	10	−2	−20
75–79	9	−3	−27
70–74	11	−4	−44
65–69	5	−5	−25
60–64	4	−6	−24
55–59	1	−7	− 7
			− 159
Total (N) . . .	107		− 53 $= \Sigma fd$

[1] This table is based on the same data as Table 11.

to a distribution similar to the one above, it involves rather laborious arithmetic computation.

Computing the mean by the short method. The last method we shall consider is the least laborious. For illustration, we may again make use of the data in Table 17. The steps in this method may be summarized as follows:

1. Tabulate the data in the form of a frequency distribution.
2. Select an assumed or *guessed* mean. This is done by taking the midpoint of an interval near the middle of the distribution. In Table 17 this would be the class interval 90–94. The midpoint is 92.5.
3. Note the *deviation* of each class interval above and below the interval in which the assumed mean lies. The deviations above the assumed mean are considered positive, and those below are considered negative. Thus in Table 17 there are seven intervals above and seven intervals below the interval containing the assumed mean.
4. Multiply algebraically each deviation (d) by its corresponding frequency (f). The products make up the column (fd). (Be sure to take the signs of the numbers into account.)
5. Find the algebraic sum of the (fd) column. This total is expressed in terms of the formula as Σfd, and in Table 17 is -53. If this sum should be zero, the assumed mean is also the true mean. If there is a positive or a negative remainder, the assumed mean has been over- or underestimated. In this case the correction (c) is made by dividing Σfd by the total number of measures (N), which in the above case is 107.
6. Since we have neglected the class interval in noting the d's and consequently the fd's, we must multiply the correction obtained (c) by the number of units in the class interval, which in this case is 5.
7. The correction is added algebraically to the guessed mean.

Carrying out these directions in connection with the data in Table 17, we obtain the following:

Estimated mean $= 92.50$

Correction $= \dfrac{\Sigma fd}{N} \times \text{Interval} = \dfrac{-53}{107} \times 5 = -2.48$

Mean $= 90.02$

It will be seen that the means obtained by the last two methods are identical. However, the means thus obtained are not necessarily identical with the mean which would be obtained by computing it in the usual manner from ungrouped data. This is because of the assumption that the midpoints of the class intervals in grouped data are the accurate averages of the frequencies grouped opposite them. In the long run this may be approximately true, but it is only by chance that they are the exact averages of the midpoints.

The median. The median is that point in a series of measures on each side of which one half of the measures lie when they are arranged in order of their magnitude. In other words, it is the point in a series or distribution that divides the series into halves so that one half includes the larger measures and the other half the smaller measures. In a simple series it may be located by counting to the halfway mark from either extreme of the distribution. For example, if 21 boys are arranged according to height from the tallest to the shortest, No. 11, counting from either end of the distribution, is the boy of median height. Ten boys are taller and ten shorter than he. If the number of cases given is even, it is customary to take the average of the two middle numbers as the median.[1]

---

[1] Strictly speaking, the measure of central tendency described in this illustration is the midscore and not the median. Many writers, however, do not distinguish between the two, although the median is computed from a frequency distribution.

This method for finding the median, although it has the advantage of simplicity, is rather cumbersome and prolonged when many cases are involved. In such instances time is saved by computing the median from a frequency distribution. The method is illustrated by the data in Tables 16 and 17, while the procedure is shown in Table 18.

TABLE 18

ILLUSTRATING METHOD OF COMPUTING THE MEDIAN FROM A
FREQUENCY TABLE

CLASS INTERVAL	f		
125–129	1		
120–124	2		
115–119	3		
110–114	3		
105–109	8		
100–104	11		Sum of frequencies from the highest
95–99	14	(42)	down to the chosen interval
90–94	13		
85–89	12	(52)	Sum of frequencies from the lowest up
80–84	10		to the chosen interval
75–79	9		
70–74	11		
65–69	5		
60–64	4		
55–59	1		
Total (N)	107		

Steps in computing the median. When the data are tabulated as shown in Table 18, the median is computed by the following steps:

1. Divide the total number of cases (N) by 2, because, as already stated, the same number of measures lie on each side of the median. In terms of Table 18, $\dfrac{N}{2} = 53.5$.

2. Beginning at the lower end of the frequency column, add the frequencies so as not to exceed half of (N). The approximate median is in the interval above the last figure added. In this table, 52 is the sum of frequencies not to exceed half of (N). Therefore 90 is the approximate median.

3. The correction of the approximate or chosen median is made by taking one half of (N), subtracting the number of cases obtained in locating the approximate median, and dividing by the frequencies of the interval in which the true median lies. The result is then multiplied by the number of points in the class interval. In the above illustration this is 53.5 minus 52.0 divided by 13 times 5 equals .58.

4. Add the correction to the approximate median. In the illustration this is 90.00 plus .58 equals 90.58.

These steps may be summed up concretely:

1. $\dfrac{N}{2} = \dfrac{107}{2} = 53.5$

2. Approximate median = 90.0 (Located by adding lowest frequencies (52) so as not to exceed $\dfrac{N}{2}$)

3. Correction $= \dfrac{(53.5 - 52)}{13} \times 5 = .58$

4. The correction added to the approximate median = 90.58. The true median is therefore 90.58.

This procedure may be checked by adding down from the upper end of the distribution:

1. $\dfrac{N}{2} = \dfrac{107}{2} = 53.5$ (Same as above)

2. Approximate median = 95.0 (Located by adding frequencies downward so as not to exceed $\dfrac{N}{2}$ or 53.5)

3. Correction $= \dfrac{(53.5 - 42)}{13} \times 5 = 4.42$

4. The correction subtracted from the approximate median equals 90.58. (The reason for subtracting the correction is that the correct median lies in the interval below the approximate median.)

The median obtained from the data in Table 18 is slightly different from the mean obtained from the same data in Tables 16 and 17. The measures of central tendency obtained by the three methods illustrated will not be exactly the same unless the distribution is symmetrical, as in the normal frequency distribution.

The mode. The mode, as a measure of central tendency, has been referred to in Chapter III, in which the method for finding the crude mode and its use were discussed. The true mathematical mode is computed by rather advanced mathematical procedures that are beyond the scope of this book. As has been suggested (page 48), in most work with educational data the mode is seldom used except as a rough inspectional average.

Which measure of central tendency to use. In working with standardized tests, the median is the measure of central tendency which is most commonly used, although the mean is sometimes computed for the same data. We have seen that in the frequency distribution that deviates from the normal distribution, there will be a difference between these two measures of central tendency. For this reason it is sometimes desirable to compute both measures when reporting educational data.

II. CALCULATING PERCENTILE POINTS IN THE DISTRIBUTION

The use of percentiles. It is sometimes helpful in analyzing a frequency distribution to have points for comparison

other than measures of central tendency. For example, we may wish to know whether a pupil's score, in relation to his group, places him in the highest fourth, the second fourth, etc., of his group. This also makes it possible to compare the scores of pupils obtained in tests which are stated in terms of dissimilar units. Thus, if with a score of 73 in a reading test and a score of 27 in an arithmetic test, a pupil is rated in the highest fourth of his class in both subjects, we have a rough but practical way of equating the two scores.

The first and third quartiles. The first and third quartiles are also referred to as the lower and upper quartiles, or as Q_1 and Q_3. They are similar to the median, which may be regarded as Q_2, in that they are all points that mark off a definite proportion of measures from the remaining measures in the distribution. The median, as has been shown, is the point chosen which separates the best or highest half from the lowest half. In the same manner, the lower quartile, Q_1, separates the lowest fourth from the highest three fourths; and the upper quartile, Q_3, separates the highest fourth from the lowest three fourths.

The method for computing Q_1 and Q_3, with the exception of the first step, is the same as that used for the median. Turning back to page 58, we see that in the first step we take half of the total number of cases (represented by $\frac{N}{2}$ or .50 N). In the first step for computing Q_1 we find the value of $\frac{N}{4}$ or .25 N, because by definition one fourth (25%) of the measures fall below Q_1. Similarly, as the first step for computing Q_3 we find the value of $\frac{3N}{4}$ or .75 N, because by definition three fourths (75%) of the measures fall below Q_3. The procedure for computing the quartiles Q_1 and Q_3 for the data in Table 18 may be illustrated as follows:

Steps in computing Q_1:

1. $\dfrac{N}{4}$ or $.25\,N = \dfrac{107}{4} = .25(107) = 26.75$

2. Approximate $Q_1 = 75$ (Located by adding the lowest frequencies (21) so as not to exceed Q_1. The next 9 would have taken us beyond the required number.)

3. Correction $= \dfrac{5.75}{9} \times 5 = 3.2$

4. Add correction to approximate $Q_1 = 75 + 3.2 = 78.2$

5. Therefore the correct Q_1 is 78.2

Steps in computing Q_3:

1. $\dfrac{3\,N}{4}$ or $.75\,N = \dfrac{3(107)}{4}$ or $.75(107) = 80.25$

2. Approximate $Q_3 = 100$ (Located by adding the lowest frequencies (79) so as not to exceed Q_3. The next 11 frequencies would have taken us beyond the required number.)

3. Correction $= \dfrac{80.25 - 79}{11} \times 5 = .6$

4. Add correction to approximate $Q_3 = 100 + .6 = 100.6$

5. Therefore the correct Q_3 is 100.6.

We have thus located three quarter points in the distribution shown in Table 18. Thus Q_1 is 78.2, Q_2 is 90.58, and Q_3 is 100.6. From these quarter points we can determine in which fourth of the distribution a given measure falls. We may also note that those measures between Q_1 and Q_3 include the middle 50 per cent, because the highest 25 per cent are above Q_3 and the lowest 25 per cent are below Q_1. It is sometimes helpful, when standard tests have been given, to regard the middle 50 per cent as the normal or average group in a class, the highest fourth as the superior group, and the lowest fourth as the slow group.

Other percentiles. Sometimes it is desirable to determine other percentile points in a distribution besides the ones discussed above. For example, we may wish to determine in which tenth, or decile, a pupil's score falls. If such percentile points should be desired, they may be found by the method for computing quartile points. Suppose, for example, we wish to determine the 90 percentile, that measure which demarks the highest 10 per cent from the lowest 90 per cent of a distribution; we may then proceed as follows (using the data in Table 18):

Steps in computing the 90 percentile:

1. $.90 \ N = .90(107) = 96.3$
2. Approximate 90 percentile $= 105$ (Located by adding the lowest frequencies (90) so as not to exceed the 90 percentile. Including the next 8 would have taken us beyond the required number.)
3. $\dfrac{96.3 - 90.0}{8} \times 5 = 3.9 =$ Correction
4. Add correction to the approximate 90 percentile $= 105 + 3.9 = 108.9$
5. Therefore the 90 percentile is 108.9.

By similar procedures any other desired percentile may be obtained. Aside from the quartile points, the percentile points most commonly computed are the deciles (tenths), and the quintiles (fifths). Using percentiles as norms to interpret and compare scores in a standardized test is illustrated in Table 19, which is taken from the Manual of Directions for the Terman Group Test of Mental Ability.[1] Obviously such a table is an aid and gives added meaning to the scores.

[1] Manual of Directions for Terman Group Test of Mental Ability, Table 1, page 9. World Book Company, Yonkers-on-Hudson, New York; 1926.

TABLE 19

Illustrating Use of the Percentile Scores According to Grade in the Terman Group Test of Mental Ability

Grade...................................		7	8	9	10	11	12
	1 per cent equal or exceed	147	170	181	194	203	207
	2½ " " " " "	134	159	172	185	196	200
	5 " " " " "	122	148	164	177	189	194
	10 " " " " "	109	135	151	166	180	185
	15 " " " " "	100	126	142	159	174	179
Upper	20 " " " " "	93	118	135	152	168	174
Quartile	25 " " " " "	88	112	128	147	163	169
	30 " " " " "	83	107	123	141	158	165
	40 " " " "	75	97	113	131	147	156
Median	50 " " " " "	68	89	104	122	138	147
	60 " " " " "	61	81	95	113	128	138
Lower	70 " " " " "	54	73	86	103	118	128
Quartile	75 " " " " "	51	69	81	98	112	122
	80 " " " " "	47	64	76	92	105	115
	85 " " " " "	43	58	71	86	99	109
	90 " " " " "	38	52	63	79	90	100
	95 " " " " "	31	43	53	67	77	86
	97½ " " " " "	25	36	44	58	66	74
	99 " " " " "	20	30	35	48	55	63
Number of cases for each grade . . .		5582	9087	10881	6730	4206	4886

Total number of cases 41,241

III. MEASURES OF VARIABILITY

The inadequacy of measures of central tendency for describing a frequency distribution. In spite of the great value of measures of central tendency in describing frequency distributions, they are not always adequate for this purpose. Such distributions may differ from each other in central tendency and also in the variability of the measures from this average. A difference in variability is illustrated in the two distributions in Table 20. It will be seen that the

measures in Column I range from 55 to 140, a difference of 85 points, while those in Column II range from 65 to 130, a difference of 65 points. It is thus clear that, although these two distributions have very nearly the same arithmetic mean, they differ in variability. The measures in II are grouped more closely about the mean than are those in I. In other words, the latter group is not so homogeneous as the former. In this section we shall discuss the measures of variability that are more commonly used.

TABLE 20

ILLUSTRATING DIFFERENCE IN VARIABILITY IN TWO DISTRIBUTIONS WITH APPROXIMATELY THE SAME MEAN

CLASS INTERVAL	I	II
135–139	2	
130–134		
125–129	7	1
120–124	8	2
115–119	11	4
110–114	7	4
105–109	11	5
100–104	12	7
95–99	11	11
90–94	11	13
85–89	10	11
80–84	10	3
75–79	11	2
70–74	7	1
65–69	7	1
60–64		
55–59	3	
Totals	128	65
Means	97.15	97.35

The range. The total range is the simplest measure of variability to obtain. It is determined by the difference

between the lowest and the highest measure in a distribution. Thus the range for Column I in Table 20 is 55 to 140, or 85, while that for Column II is 65 to 130, or 65. The range as a method of measurement is likely to be very unreliable, however, because it is determined by the extreme measures. In Table 20, for example, if we eliminate the extreme measures in Column I, the range is the same as that of Column II. For this reason the range may be determined by one measure at each extreme of the distribution, with the result that it is of value only as a very rough measure of variability.

The quartile deviation. The quartile deviation is computed from the formula: $\dfrac{Q_3 - Q_1}{2}$. When Q_3 and Q_1 have been determined, it becomes an easy task to calculate the quartile deviation. Thus, for the data in Table 18 the quartile deviation is: 11.2. Strictly speaking, the quartile deviation is not a deviation at all, because it is not determined with reference to a measure of central tendency, but is found by locating Q_3 and Q_1. However, it has been widely used in ordinary work with educational statistics because it may be easily computed and understood.

The standard deviation. The symbol used to designate the standard deviation is the Greek letter sigma (σ). It is usually computed from the mean, although it may be computed from other measures of central tendency. For a simple series the formula is: $\sigma = \sqrt{\dfrac{\Sigma d^2}{N}}.$ For a frequency distribution the formula becomes: $\sigma = \left(\sqrt{\dfrac{\Sigma f d^2}{N}}\right) i.$ If the standard deviation is computed from an assumed mean, as is usually the case, the necessary correction results in the formula: $\sigma = \left(\sqrt{\dfrac{f d^2}{N} - c^2}\right) i.$

TABLE 21

ILLUSTRATING COMPUTATION OF THE STANDARD DEVIATION FROM A SIMPLE
SERIES SHOWING THE DISTRIBUTION OF THE HEIGHTS OF A GROUP OF
BOYS

HEIGHT IN INCHES	d	d^2
62	6	36
61	5	25
60	4	16
59	3	9
58	2	4
57	1	1
56	0	0
55	— 1	1
54	— 2	4
53	— 3	9
52	— 4	16
51	— 5	25
50	— 6	36
Mean = 57		$182 = \Sigma d^2$

The notation used in these formulas may be explained as follows:

σ is the standard deviation (sigma) of the measures or scores tabulated in a distribution.

N is the number of cases in the distribution.

f designates frequency.

d designates the deviation of a class interval from an assumed mean.

c is the correction which must be computed in calculating the mean from a frequency distribution. This correction is also used in determining the standard deviation from a frequency distribution in which class intervals are used.

Σ designates summation.

i represents the class interval of the distribution.

The computation of the standard deviation from a simple series may be illustrated with the data in Table 21. Substituting the values obtained in the first formula on page 66, we have: $\sigma = \sqrt{\frac{182}{13}} = \sqrt{14} = 3.742$.

We may now proceed to calculate the standard deviation from a frequency distribution. For this purpose, we may repeat the data of Table 20, Column I, in the following table:

TABLE 22

ILLUSTRATING CALCULATION OF STANDARD DEVIATION FROM A FREQUENCY DISTRIBUTION

CLASS INTERVAL	f	d	fd	fd^2
135–139	2	8	16	128
130–134		7		
125–129	7	6	42	252
120–124	8	5	40	200
115–119	11	4	44	176
110–114	7	3	21	63
105–109	11	2	22	44
100–104	12	1	12	12
95–99	11	0		
90–94	11	− 1	− 11	11
85–89	10	− 2	− 20	40
80–84	10	− 3	− 30	90
75–79	11	− 4	− 44	176
70–74	7	− 5	− 35	175
65–69	7	− 6	− 42	252
60–64		− 7		
55–59	3	− 8	−24	192
Total (N) . .	128		− 9	$1811 = \Sigma fd^2$

The column headings in Table 22 are the same as those in Table 17 for calculating the mean, with the exception of the last column, which is obtained by multiplying the values in the fd column by the values in the d column. The values obtained are:

$N = 128$ (sum of the column of frequencies)

$\Sigma fd = 1811$ (sum of the fd^2 column)

$$c = \frac{-9}{N} = \frac{-9}{128} = -.07$$

$i = 5$ (number of points in each interval)

Substituting these values in our third formula on page 66, we obtain:

$$\sigma = \left(\sqrt{\frac{1811}{128} - .0049} \right) \times 5$$
$$= (\sqrt{14.14}) \times 5$$
$$= 18.8$$

A number of advantages of the standard deviation as a measure of variability have been listed by Rugg [1] as follows:

1. It is numerically defined.
2. It is based on all the measures.
3. It is easily calculated.
4. It is susceptible of algebraic treatment.
5. It can be shown by the theory of errors and sampling that it is the measure of variability least affected by fluctuations of sampling.
6. Its computation aids the determination of the Pearson coefficient of correlation.
7. It is convenient because of the necessity of obtaining a measure which will vary with the variability of distribution, and squaring deviations is the simplest method of eliminating signs.
8. It bears a convenient relationship to the normal or probability curve. . . .

On page 70 it will be seen that the standard deviation is used in calculating the probable error $(P.E.)$; and in later chapters we shall see that it is used to compute such derived scores as T-scores from the point scores of a test.

---

[1] H. O. Rugg, *Statistical Methods Applied to Education*, page 173. Houghton Mifflin Company, Boston; 1917. By permission of the publishers.

The probable error. The probable error (*P.E.*) is sometimes used for the quartile deviation because in a normal distribution it is numerically identical with that deviation. It has therefore been used as a convenient measure of variability because it may be determined with ease if the standard deviation is already known. It may be obtained by multiplying the standard deviation by .6745; that is, $P.E. = .6745\,\sigma$. As is also true of the quartile deviation, the probable error is so related to the normal frequency surface that when it is measured off on the base line on each side of the mean and perpendiculars are erected, one half of the frequency surface will be included. If 2 *P.E.* is laid off in this way, about 80 per cent of the surface will be included; while if 3 *P.E.* is so measured, about 95 per cent will be included. Similar calculations have been made for 4 *P.E.*, 5 *P.E.*, etc., and are arranged [1] in tabular form for convenient use.

The facts just stated are useful when the probable error is used to indicate the degree of error in such group measures as those discussed in Section V (pages 81 to 83). For example, when a given mean has been calculated, we can say that the chances are even (1 to 1) that the true mean lies within the limits of 1 *P.E.* added to and subtracted from the obtained mean. In the same manner, we can say that the chances are approximately 80 to 20 (4 to 1) that the true mean lies within 2 *P.E.* added to and subtracted from the obtained mean. Analogous statements can be made for 3 *P.E.*, 4 *P.E.*, etc. These matters will be discussed more fully in Section V.

IV.   CORRELATION

The meaning of correlation. In the statistical sense correlation is the study of paired facts. For example, a teacher

[1] See C. B. Davenport, *Statistical Methods*, page 14. John Wiley and Sons, New York; 1904.

may wish to compare marks or scores obtained by her pupils in arithmetic with those obtained in history.   To be reliable, such comparisons should be based on marks or scores that are trustworthy.   Comparisons based on unreliable marks, such as those discussed in Chapter II, would be very misleading. For this reason our first concern is to obtain scores based upon objective and reliable tests.   Let us suppose that this has been done and the results shown in Table 23 for a class of ten seventh-grade pupils are obtained.

TABLE 23

SHOWING POSSIBLE RESULTS OF AN OBJECTIVE TEST GIVEN TO A CLASS OF TEN SEVENTH-GRADE PUPILS

PUPIL	ARITHMETIC		HISTORY	
	Score	Rank	Score	Rank
A	34	1	76	1
B	33	2	72	2
C	30	3	66	3
D	28	4	63	4
E	25	5	60	5
F	22	6	58	6
G	21	7	54	7
H	20	8	50	8
I	18	9	48	9
J	16	10	44	10

It is evident that Pupil A is first in both subjects, that Pupil B is second, that Pupil C is third, etc.   This is shown by placing each pupil's rank after his score.   It is evident also that the table illustrates a perfect positive relationship.

This relationship may also be represented graphically in the form of a double distribution table such as is shown in Table 24.

It will be seen that when tabulated in this way (with the

TABLE 24

ILLUSTRATING PERFECT POSITIVE RELATIONSHIP

		RANK IN HISTORY									
		1	2	3	4	5	6	7	8	9	10
RANK IN ARITHMETIC	10										1
	9									1	
	8								1		
	7							1			
	6						1				
	5					1					
	4				1						
	3			1							
	2		1								
	1	1									

vertical class intervals in the order of the highest at the top to the lowest at the bottom, and with the horizontal class intervals in the order of the lowest at the left to the highest at the right), perfect positive correlation presents a straight-line relationship extending from the upper right- to the lower left-hand corner of the table. Deviation from such correlation would result in a scattering of cases on both sides of this line, depending upon the amount of deviation.

Let us now suppose that the application of our tests in arithmetic and history has given us a different result such as illustrated in Table 25.

It is evident that, in this illustration, the situation is exactly opposite from the one shown in the preceding illustration. Here Pupil A ranks first in arithmetic but tenth in

history, Pupil B ranks second in arithmetic but ninth in history, etc. These distributions illustrate a perfect negative relationship.

When represented in the form of a double distribution table in the manner described for Table 24, the above table appears as in Table 26. From this table, illustrating the exact opposite of the relationship presented in Table 24, we see that perfect negative correlation presents a straight-line relationship extending from the upper left- to the lower right-hand corner of the table.

Calculation of the coefficient of correlation. Such tables as 24 and 26 are useful in making rough estimates of the degree of relationship that exists between two traits for which we have adequate measurements. It is often desirable, however, to have a more precise statement of the amount of relationship than can be obtained by mere inspection. For this purpose a formula has been devised for computing

TABLE 25

SHOWING POSSIBLE RESULTS OF AN OBJECTIVE TEST GIVEN TO A CLASS OF TEN SEVENTH-GRADE PUPILS

PUPIL	ARITHMETIC		HISTORY	
	Score	Rank	Score	Rank
A	34	1	44	10
B	33	2	48	9
C	30	3	50	8
D	28	4	54	7
E	25	5	58	6
F	22	6	60	5
G	21	7	63	4
H	20	8	66	3
I	18	9	72	2
J	16	10	76	1

TABLE 26

ILLUSTRATING PERFECT NEGATIVE RELATIONSHIP

		RANK IN HISTORY									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	10	1									
	9		1								
	8			1							
RANK IN ARITHMETIC	**7**				1						
	6					1					
	5						1				
	4							1			
	3								1		
	2									1	
	1										1

the coefficient of correlation [1] from such tables as 24 and 26. This formula enables us to state the amount of relationship in exact numerical terms that range from -1.00 for perfect negative correlation to $+1.00$ for perfect positive correlation. The coefficients of correlation, when they are computed between scores obtained by a class in any two elementary school subjects, are usually between 0.3 and 0.9, depending on the subjects and the type of tests used. In other words, such correlations are positive and indicate that pupils tend to do as well in one school subject as in another.[2]

[1] The letter r is used as a symbol to designate the coefficient of correlation.
[2] The detailed steps for calculating the coefficient of correlation, which are rather difficult for beginners, have been placed in the Appendix. At the discretion of the instructor the class may turn to a study of this part of the book.

Some uses of the coefficient of correlation. There are many statistical considerations related to the coefficient of correlation that are too advanced for discussion in this book. As a matter of fact the classroom teacher rarely has occasion to make the actual computations of correlation coefficients. Nevertheless, it is important that she should have a general understanding of the technique and its application. In the first place, it will be impossible for her to understand modern educational articles and books without an understanding of the correlation technique, since it is used so often in the reports of educational experiments and investigations by research workers. In the second place, the coefficient of correlation is one of our most useful tools for determining the validity and the reliability of standardized tests.

Validity. Validity has been defined as the degree to which a test measures what it claims to measure. Thus a test that claims to measure general information in American history is not valid if it is made up of ambiguous or unimportant questions. To be a valid measure, it should be based upon an adequate sampling of test items taken from important topics of the subject. In other words, just as the information in a good course of study in American history should be founded on socially useful material adapted to the grade or age of the pupils concerned, so a good standardized test in the subject should give a valid measure of the progress which a pupil has made after his study of the subject. This presupposes, among other things, that the items of the test have been carefully selected experimentally, both as to importance and as to the degree of difficulty for the pupils concerned. One way to judge the validity of a new test is to give it to a suitable group of pupils who have previously taken a test of known validity, and then to compute the coefficient of correlation between the tests. If the correlation between the new test and the old test is high, we have important evidence of

the validity of the new test. If a validated test of the same type is not available, the new test may be correlated against teachers' judgments or the judgments of experts.

Reliability. By reliability is meant the accuracy with which a test measures what it does measure. Thus a test which gives radically different results on successive applications with the same group may be called unreliable. A test must be reliable to be valid. However, reliability does not guarantee validity. For example, a test might be intended as a test of information in American history, and yet the peculiar wording of the exercises may make it a test of general intelligence. Furthermore, it might be a very accurate measure of intelligence.

The reliability of a test may be determined in several ways. A common method is to compute the coefficient of correlation between two forms of the same test, both of which have been given to the same group of pupils. Another method, used when there is only one form of the test, is to compute separately the scores obtained on the odd-numbered items and the scores on the even-numbered items. Thus the test is divided into halves, the scores on which may be correlated. By a comparatively simple formula the correlation between these halves is then corrected for length of test in order to determine what the correlation for the entire test would be. A third method is to repeat a test after an interval of time, using the same form, and then compute the coefficient of correlation between the scores from these two applications. This method is seldom satisfactory, because familiarity with the test items may influence the correlation. A coefficient of correlation computed by any of the three methods just described is called a reliability coefficient. Such coefficients should always be included in the information given the test user, together with the descriptive material pertaining to the administration and interpretation of any given test. Relia-

bility coefficients, in addition to similar information, enable one to discriminate between tests that vary in reliability.

A reliability coefficient by itself, however, does not permit a complete evaluation of the reliability of a test. The explanation of this statement would take us beyond the scope of this book, but proof may be found in several of the reference books listed at the end of the chapter. Among the other necessary or helpful facts which should be given with the reliability coefficient are the following:

1. A description of the group used in determining the reliability coefficient, including range of grades or ages and other information from which one can judge whether or not the group is representative.
2. The standard deviations of the tests used in computing the reliability coefficients. These must be known because the significance of the size of a reliability coefficient depends in part on the range of talent from which it is computed, and the standard deviation is the necessary measure of the range of talent.
3. A statement of the method used in computing the reliability coefficient.
4. The number of cases used.
5. The average scores (mean or median) of the tests used.

The facts listed above will not only aid in determining the reliability of a test but also make possible calculations that are sometimes necessary in connection with the evaluation of tests. While it is not likely that the classroom teacher will be called upon to make calculations involving the facts listed, she should nevertheless be aware of their relation to the reliability coefficient of a test. She should form the habit of looking for them in the descriptive material accompanying the tests she plans to use, just as a careful housekeeper examines the labels required by pure-food laws.

Satisfactory reliability of a test depends to a considerable extent upon the purpose for which the test is to be used. It should also be kept in mind that the range of talent influences the size of the reliability coefficient, other things being equal. In other words a reliability coefficient based on testing of two or more grades will be larger than one based on testing of only one grade. Ruch and Stoddard [1] suggest the following interpretation of reliability coefficients as a rough guide. They state that " these figures are suggestive only and imply that the reliability has been computed on average or typical classes of a sufficient size to provide a stable sample."

RELIABILITY COEFFICIENT	INTERPRETATION OR SIGNIFICANCE
0.95 to 0.99	Very high; rarely found among present tests.
0.90 to 0.94	High; equaled by a few of the best tests.
0.80 to 0.89	Fairly high; fairly adequate for individual measurement.
0.70 to 0.79	Rather low; adequate for group measurement but not very satisfactory for individual measurement.
Below 0.70	Low; entirely inadequate for individual measurement although useful for group averages and school surveys.

Kelley [2] groups all testing under six purposes and suggests the reliabilities requisite for each purpose:

1. The measurement of the general group (grade or school) accomplishment and an estimate of the probable future general group success in school work.

2. The measurement of a school group in some specific subject and an estimate of the future group promise in the same or a closely related subject.

3. The measurement of the relative differences in achievement of the group in two or more scholastic lines and an estimate of the significance of such differences.

[1] G. M. Ruch and George D. Stoddard, *Tests and Measurements in High School Instruction*, pages 55–56. World Book Company, Yonkers-on-Hudson, New York; 1927.
[2] T. L. Kelley, *Interpretation of Educational Measurements*, pages 28–29. World Book Company, Yonkers-on-Hudson, New York; 1927.

4. The measurement of the past general scholastic success and the future promise of an individual.

5. The measurement of the success of an individual in a specific school subject and an estimate of his future promise in the same or a closely related subject.

6. The measurement of differences in the individual of abilities and accomplishments in several scholastic lines and an estimate of the probability of persistence of differences, of the sort revealed, in future school work or vocation.

With reference to the purposes listed, Kelley says, " The minimal satisfactory reliabilities, as measured by a reliability coefficient determined from the pupils in a single school grade, of tests serving these six purposes are as follows: .50, .50, .90, .94, .94, and .98, respectively." It appears from the foregoing that Kelley insists on higher reliability coefficients for individual measurement than do Ruch and Stoddard, but Ruch and Stoddard set higher requirements for group measurement than those of Kelley. The authorities cited above set requirements for satisfactory reliability coefficients for measuring both the individual and the group higher than did the earlier authorities on the subject, such as Rugg.[1] Because only a few of the best tests have reliability coefficients as high as 0.90, Kelley's requirement of a reliability coefficient as high as 0.98 [2] for his sixth purpose would practically preclude the use of standard tests for the measurement of individual differences. This would mean a return to subjective estimates which, as we have seen, are of little reliability. We must therefore forego the attempt to measure for the sixth purpose listed by Kelley, or be content with tests of the reliability suggested by Ruch and Stoddard.

[1] Harold O. Rugg, *Statistical Methods Applied to Education*, pages 256–257. Houghton Mifflin Company, Boston; 1917.

[2] The Stanford Achievement Test, which is discussed in Chapter VIII, does approximately meet this requirement when the total score is used. Ultimately we may expect that standardized tests of high reliability will become more numerous.

V. THE RELIABILITY OF MEASURES

The reliability of group measures. In order to obtain a true measure of a group — for example, the true average — one must obtain accurate measurements of each individual in the group. Thus in order to find the true average height of eighth-grade boys in the United States, it would be necessary to know the height of each boy in this grade. In this manner one might also obtain true measures of variability; for example, the standard deviation. The true coefficient of correlation likewise could be obtained between any two traits for which measurements are available. In computing group measures, however, it is practically impossible to obtain measurements of all the individuals in a group. Even if this were possible, the amount of time required would usually be prohibitive. The practicable procedure is to obtain measures of a part of the group so selected as to be representative or typical of the whole. Group measures obtained in this way will approximate rather closely the results that would have been obtained by including the " total population," provided care has been taken to insure a random sampling of the whole group.

The meaning of a random sample. The principle of random sampling is well illustrated by the use of " straw votes " in an election. When care is exercised to secure random sampling, straw votes may predict with surprising accuracy the final outcome of an election. This may be illustrated by the nation-wide presidential poll conducted by the *Literary Digest* [1] during the campaign of 1928. The following compilation shows how closely and consistently the final official vote was predicted by relatively small samplings before the election :

[1] See *Literary Digest* for October 6, 13, 20, 27, and November 3, 1928; also for January 5, 1929. Figures are reproduced here by permission of Funk & Wagnalls Company, publishers.

TABLE 27

SHOWING RESULTS OF STRAW VOTE TAKEN BY THE *Literary Digest* DURING
THE PRESIDENTIAL CAMPAIGN OF 1928

DATE	NUMBER OF VOTES CAST			PER CENT	
	Hoover	Smith	Total	Hoover	Smith
6 October	514,397	231,061	745,458	69.05	30.95
13 October	1,201,869	688,829	1,890,698	63.56	36.44
20 October	1,593,436	910,234	2,503,670	63.64	36.36
27 October	1,717,041	971,356	2,688,397	63.87	36.13
3 November	1,750,584	987,795	2,738,379	64.21	35.79
Final Vote	21,429,109	15,005,497	36,434,606	58.81	41.19

We may note that, except during the first week, the per-
centages for the two candidates remain very nearly the same
throughout. We may note also that they consistently
underestimate the final vote cast for Smith. This indicates
that the *Literary Digest* poll did not adequately sample the
parts of the electorate more favorably disposed to Smith
than to Hoover.

The reliability of measures in terms of the probable error.
Certain formulas have been devised for determining the error
due to sampling found in group measures, such as the mean,
the standard deviation, and the coefficient of correlation.
We may illustrate the use of these formulas in connection
with the measures just listed. Suppose, for example, we
desire to determine the reliability of the mean for the first
distribution in Table 20 (page 65). We may do this by the
application of the formula:

$$P.E._M = \frac{.6745\ \sigma}{\sqrt{N}}$$

The obtained mean for this distribution is 97.15, the standard
deviation (σ) is 18.8, and N is 128. Substituting these

values in our formula, we find that the probable error of the obtained mean $(P.E._M)$ is 1.12. This indicates that the chances are even that the obtained mean does not differ from the true mean by more than 1.12. In other words, the chances are even that the true mean lies between 96.03 and 98.27. The chances are about four to one that the obtained mean does not differ from the true mean more than 2.24 or that it is between 94.91 and 99.39. It will be seen by inspection of the formula given on page 81 that the magnitude of $P.E._M$ depends upon the magnitudes of N and σ. That is, $P.E._M$ decreases as N increases or as σ decreases. This is equivalent to saying that the true mean is more nearly approximated as the number of cases (N) is increased or as the variability (σ) of the distribution is decreased.

$P.E._M$ is especially important in determining the reliability of norms; for example, a norm for a given grade or age. When we compare scores with the norm, we need to know how trustworthy or reliable it is. If norms are based upon the measurement of pupils who are not typical for their age or grade, or if they are based upon too small a sampling, they may be very misleading as standards of comparison. On the other hand, if reasonable care is taken to derive norms from test scores of representative pupils, a few thousand cases will suffice to establish reliability.

Just as an obtained mean may differ from the true mean, so the obtained standard deviation of a distribution may differ from the true standard deviation. The reliability of an obtained standard deviation may be determined very much in the same manner as that of an obtained mean. The formula is:

$$P.E._\sigma = \frac{.6745\ \sigma}{\sqrt{2N}}$$

We may illustrate with the data in Table 20, in which σ is 18.8 and N is 128. Substituting these values, we obtain:

$$\frac{.6745(18.8)}{\sqrt{2(128)}} = \frac{12.68}{16.00} = .79$$

We may say that the chances are even that the true standard deviation is 18.8 $\pm$ .79, or that it lies between 18.01 and 19.59. We may also say that the chances are about four to one that the true standard deviation is 18.8 $\pm$ 1.58, or that it lies between 17.22 and 20.38.

The probable error of a score. The reliability coefficient, as we have seen in Section IV (page 76), gives us a notion of the reliability of the test as a whole but tells us very little about the reliability of any given score. Yet, to the class-room teacher the important item is the individual score. She must know, for example, how much she can rely upon a score of, say 80, obtained by John Jones. The formula for the probable error of such a score is: $P.E._{(score)}$ = .6745 $\sigma\sqrt{1 - r_{12}}$. In this formula the standard deviation (σ) ordinarily represents the average standard deviation of the two distributions from which the reliability coefficient has been calculated, and r_{12} represents the reliability coefficient.

Let us now suppose that a pupil has obtained a score of 80 in a test for which the standard deviation is 15.00 and r_{12} is .90. Substituting these data in our formula, we obtain:

$$P.E._{(score)} = .6745(15)\sqrt{1 - .81}$$
$$= 10.12\sqrt{.19}$$
$$= 10.12 \times .43 = 4.35$$

We may now say that the chances are even that the true score in this case is 80 $\pm$ 4.35, or that it lies between 75.65 and 84.35. We may also say that the chances are about four to one that the true score is 80 $\pm$ 8.70, or that it lies between 71.30 and 88.70.

1. Find the mean, median, and mode for the following distributions:

CLASS INTERVAL	I	II	III
135–139	2	1	
130–134	2	1	2
125–129	11	5	5
120–124	16	5	7
115–119	12	7	8
110–114	12	6	11
105–109	24	4	7
100–104	18	3	11
95–99	20	11	12
90–94	32	18	11
85–89	16	12	11
80–84	18	10	10
75–79	22	7	10
70–74	14	9	11
65–69	22	6	3
60–64	15	1	9
55–59	10	5	5
50–54	7		
45–49	5	1	4
40–44	4		1
35–39		1	3
30–34			
25–29	2		
20–24		1	
Totals　.　.　.	284	114	141

2. For each of the above distributions calculate Q_1, Q_3, the 10 percentile, and the 90 percentile.

3. For each of the above distributions calculate the quartile deviation, the standard deviation, and the probable error.

4. For each of the above distributions calculate $P.E._{\cdot M}$.

5. For each of the above distributions calculate $P.E._{\cdot \sigma}$.

6. Three pupils made scores in an intelligence test of 109, 135,

and 149 respectively. In this test the standard deviation is 13 and the reliability coefficient (r_{12}) is .85. Using these data, calculate $P.E._{(score)}$ for each pupil.

References

BROWN, WILLIAM, and THOMSON, GODFREY H. *The Essentials of Mental Measurement.* The Macmillan Company, New York; 1921.

BUCKINGHAM, B. R. *Research for Teachers,* Chapter II. Silver, Burdett & Co., New York; 1926.

—— "Random Sampling in Supervision." *Journal of Educational Research,* Vol. V (May, 1922), pages 428–430.

—— "The Use of the Median in Supervision." *Journal of Educational Research,* Vol. V (February, 1922), pages 154–157.

FRANZEN, RAYMOND. "Attempts at Test Validation." *Journal of Educational Research,* Vol. VI (September, 1922), pages 145–158.

GARRETT, HENRY L. *Statistics in Psychology and Education.* Longmans, Green & Co., New York; 1926.

GREGORY, C. A. *Fundamentals of Educational Measurement,* Chapters X, XI, and XII. D. Appleton & Co., New York; 1922.

KELLEY, T. L. *Statistical Method.* The Macmillan Company, New York; 1923.

McCALL, WILLIAM A. *How to Measure in Education,* Chapters VII, XI, XIV, XV, XVI, and XVII. The Macmillan Company, New York; 1922.

MONROE, W. S. *The Theory of Educational Measurements,* Chapters IX, XII, and XIII. Houghton Mifflin Company, Boston; 1923.

ODELL, C. W. *Educational Statistics.* The Century Company, New York; 1925.

OTIS, ARTHUR S. *Statistical Method in Educational Measurement.* World Book Company, Yonkers-on-Hudson, New York; 1925.

RUCH, G. M. "Minimum Essentials in Reporting Data on Standard Tests." *Journal of Educational Research,* Vol. XII (December, 1925), pages 349–358.

—— and STODDARD, G. D. *Tests and Measurements in High School Instruction,* Chapters XVII, XVIII, XIX, and XX. World Book Company, Yonkers-on-Hudson, New York; 1927.

RUGG, H. O. *Statistical Methods Applied to Education,* Chapters V and IX. Houghton Mifflin Company, Boston; 1917.

THURSTONE, L. L. *The Fundamentals of Statistics.* The Macmillan Company, New York; 1925.

TRABUE, M. R. *Measuring Results in Education,* Chapters IX and XVII. American Book Company, New York; 1924.

CHAPTER FIVE

THE NATURE OF INTELLIGENCE AND ITS MEASUREMENT BY INDIVIDUAL TESTS

Popular concepts of intelligence. Concepts of general intelligence and of special talents and aptitudes appear to have been present in the popular mind long before the present scientific interest in these matters. Popular recognition of individual differences in mental traits and capacities and of their transmission through heredity is illustrated by such sayings as " Blood will tell," " Music runs in that family," " Like father, like son," " He's a chip off the old block." Such concepts seem to be at least as old as the recorded history of man. For example, in the Bible we find individual differences in mentality and character recognized, as in Gideon's method of selecting his army and in the parable of the talents; [1] we find also suggestions of the possibility of measuring these differences.

Plato's concept of inborn differences. In his *Republic* Plato proposes to classify all workers in three groups — the rulers or guardians, the military group or auxiliaries, and the producers, including artisans and farmers. He proposes further that the members of these vocations be selected on the basis of natural aptitudes. Plato even proposes methods of testing to determine the differences. For example, in selecting rulers, he suggests that those who might become rulers be carefully observed from early childhood in regard to their behavior and activities under complex conditions. Plato's classification of workers may seem too simple for our present complex civilization, and we may disagree with him as to the relative importance of each group. Nevertheless, he recognized as clearly as our modern psychologists the presence of important individual differences in mental traits

[1] See Judges vii. 1–7 and Matthew xxv. 14–30.

and capacities, their inborn character, and the desirability as well as the possibility of measuring them.

The scientific study of intelligence. It is perhaps natural that society should first focus scientific attention upon those individuals who deviate most from normality in intelligence, such as the genius and the feeble-minded. In these two groups the deviations from the normal or average are so great that society can hardly remain ignorant of them. This is especially true of the feeble-minded and the subnormal. Pintner [1] traces interestingly the changes in attitude from ancient times to the present. He points out the Greek and Roman practice of exposing the physically and possibly the mentally defective to the elements, the tolerant and more or less superstitious attitude accompanying the Christianity of the medieval period, the change during the Renaissance to the severe punishment of the feeble-minded and insane, and the gradual development of the scientific attitude during the modern period. As concrete examples of the scientific attitude we may cite the study of feeble-minded families, such as the Jukes and the Kallikak and, by way of contrast, of families possessing unusual talent, such as the Edwards family.

Nature versus nurture. One of the most important questions arising in the study of intelligence is that of the relative importance of heredity and environment in determining the amount of intelligence possessed by a given individual. There can be no question of the strong similarity of the individuals belonging in a given family group, such as those mentioned above. Some persons, however, do not find in the study of these families conclusive proof that the similarity is entirely due either to heredity or to environment. Numerous ingenious studies have been made that attempt

[1] R. Pintner, *Intelligence Testing*, Chapter I. Henry Holt & Co., New York; 1923.

to determine the relative importance of the two factors. Among the first is a study by Sir Francis Galton of the adopted sons of popes. Galton found that although these sons were adopted by men of great ability and were brought up in the best environment intellectually and culturally, they did not achieve the eminence of their foster fathers. On the other hand, Galton [1] found that, among families of the same social standing, the chances of becoming eminent are much greater if one is born in a family of high attainments than if one is born in a family of average attainments. Briefly, he found that 977 eminent men had 537 eminent relatives, while the same number of average men had only 4 eminent relatives.

Thorndike [2] attempted to approach the question by comparing the resemblances found between twins with those found between brothers and sisters who are not twins. He reasoned that if there was greater resemblance between twins, it would be due to nature (not nurture), since it is in the latter respect that they differ from ordinary brothers and sisters. By giving both groups a series of tests, he found that the twins resembled each other more closely. He showed further that older twins do not resemble each other more closely than younger twins, although environment has had longer opportunity to increase the resemblance. Since Thorndike's investigation several others of the same type have been made which support his main conclusions. Thus Wingfield and Sandiford,[3] in the summary of a very carefully conducted investigation, conclude by saying: " Therefore,

---

[1] Francis Galton, *Hereditary Genius: An Inquiry into Its Causes and Consequences.* The Macmillan Company, New York; 1869.

[2] E. L. Thorndike, *The Measurement of Twins* (Archives of Philosophy, Psychology, and Scientific Methods, No. 1). Columbia University, New York; 1905.

[3] A. H. Wingfield and Peter Sandiford, "Twins and Orphans," in *Journal of Educational Psychology,* Vol. XIX, pages 410–421; September, 1928.

there is an increasing degree of resemblance in general intelligence among human beings with an increasing degree of blood relationship among them. Ergo, general intelligence is an inherited trait." This does not mean that environment is not a factor. It is obvious, for example, that environment is important in determining what particular direction any given talent or capacity will take.

The studies of subnormal and supernormal families have not been accepted by everybody as conclusive evidence that intelligence is largely transmitted through heredity from one generation to the next. We may therefore present, in support of these conclusions, evidence from actual intelligence test results. These results were obtained by the writer [1] in testing high school students in a large city. Various comparisons were made between students in two high schools, one of which was largely college preparatory and classical, the other of which was largely vocational. In the tables that follow we shall call the first school " A " and the second " B." The students of both A and B were classified in five groups according to the vocations of their parents. These vocational groups were as follows:

     I.   Unskilled
    II.   Semiskilled
   III.   Skilled
   IV.   Business
    V.   Professional

While these vocations overlap to a certain extent, making it difficult to be sure that every case was properly classified, this overlapping was not so frequent that the results may not be accepted as correct so far as the general tendency which is revealed is concerned. Table 28 shows the percentage of

[1] I. N. Madsen, "The Contribution of Intelligence Tests to Vocational Guidance in High School," in *School Review,* Vol. XXX, pages 692–701; November, 1922.

students in each group for the two schools as well as the percentage and number of cases for both schools combined.

TABLE 28

SHOWING PERCENTAGE OF STUDENTS IN EACH OCCUPATIONAL GROUP IN EACH OF THE TWO HIGH SCHOOLS AND PERCENTAGE AND NUMBER FOR THE COMBINED SCHOOLS

| SCHOOL | GROUP | | | | | TOTAL |
	I	II	III	IV	V	
A	2.5	3.5	18.5	47.0	28.5	100
B	7.0	10.5	37.0	41.1	4.4	100
Both	5.7	6.9	26.6	44.5	16.3	100
Number . .	195	237	913	1527	560	3432

It will be seen from this table that most of the cases in School A are in Groups IV and V, while most of the cases in School B are in Groups II, and III, and IV. In other words, a greater percentage of the pupils in School B come from the homes where parents are engaged in the relatively less desirable vocations from the point of view of social standing and financial reward. These students were next classified (Table 29) on the basis of their scores in an intelligence test to show differences according to age, grade, and school.

We may now analyze the differences shown in Table 29. The higher scores of the younger pupils in each grade and for each school is evidence that the test used really measured intelligence. This interpretation is in accord with the logic of the situation; for the older pupils have had more life experience since they have lived longer, and have also had more school experience on the average than the younger pupils. As far as the environment is concerned, they are the favored group and should consequently obtain the higher scores if the environment determines the size of the score.

TABLE 29

COMPARISON OF INTELLIGENCE SCORES MADE BY STUDENTS OF TWO
HIGH SCHOOLS

YEAR	SCHOOL	CHRONOLOGICAL AGE										NUM-BER
		13	14	15	16	17	18	19	20	21	22	
Freshman	A	128	112	110	101	104	100	90				611
	B	91	91	86	79	72						647
Sophomore	A		135	129	122	118	123	117	70			419
	B		97	107	92	86	85	80	80			350
Junior	A			130	136	132	123	120	120	120		341
	B			97	108	106	93	80				244
Senior	A			190	150	143	140	129	125	129	115	280
	B				130	111	116	103	89			145
Age standards . .		84	87	92	96	101	107	107	109			

Again, if we take any given age, such as 15, those pupils who are further advanced obtain higher scores. In other words, of the 15-year-olds in high school, the sophomores score higher than the freshmen, the juniors higher than the sophomores, and the seniors higher than the juniors. We may reasonably conclude from this that the test really measured intelligence. We should logically expect that a pupil who has reached the senior class at 15 is more intelligent than one remaining in the freshman class at this age. Furthermore, the 15-year-olds in each of the four classes have had approximately the same environment as far as extent of life experience and home experience are concerned.

We may use somewhat the same reasoning in explaining the difference in the intelligence shown in the two schools. One might be tempted to conclude that the better home

environment afforded the students of School A, as indicated by Table 28, accounts for this disagreement. However, environment cannot account for the entire difference, for if it did the older pupils in each class in both School A and School B would have obtained the higher scores because they had been the favored ones environmentally. In other words, the difference of scores in favor of School A indicates a real superiority of intelligence over School B; and since heredity is the one other respect in which the pupils differ, it follows that the pupils have, at least in part, inherited intelligence from their ancestry.

Binet's concept of intelligence. We have already called attention in Chapter I to Binet's experiments in measuring mental traits, experiments that resulted in the first general intelligence test. It is interesting to note here that this test grew out of an attempt to determine the degrees of brightness among feeble-minded children in Paris. With this purpose Binet and Simon were given the commission to make this determination by examination of the children concerned. In this connection the first Binet-Simon scale was published and used. This fact again illustrates the statement made earlier in the chapter that scientific attention was first focused upon the individuals who deviate noticeably from the normal.

Binet departed in three important respects from his previous methods and from the methods for measuring intelligence used by contemporary psychologists: (1) He made use of age standards as a basis for comparison; (2) he endeavored to test the higher and more complex mental processes rather than the simpler and more elementary ones; (3) he attempted to test " general intelligence " as a whole, rather than to test separate elements and then combine the results. These methods will be made clear in the description of the Binet-Simon Scale in the following pages. In his concept of general intelligence, Binet emphasized three

phases of thinking: (1) the ability to take and to maintain a definite direction; (2) the capacity to make adaptations in order to attain a desired end; and (3) the power of self-criticism.

Other concepts and definitions of intelligence. A much-quoted definition is one by the German psychologist, Stern, who defines intelligence as the " general capacity of an individual consciously to adjust his thinking to new require-ments; it is general adaptability to new problems and condi-tions of life." From among the more recent definitions we may quote a few from a symposium on intelligence published in the *Journal of Educational Psychology.*[1] Since some of the definitions are rather long or are modified by accompany-ing explanations, only a few examples will be given. Among these, we may quote the following:

We may then define intellect in general as the power of good response from the point of view of truth and fact. THORNDIKE

An individual is intelligent in proportion as he is able to carry on abstract thinking. TERMAN

An individual possesses intelligence in so far as he has learned, or can learn, to adjust himself to his environment. COLVIN

Intelligence is intellect plus knowledge. HENMON

Intelligence seems to be a biological mechanism by which the effects of a complexity of stimuli are brought together and given a somewhat unified effect in behavior. PETERSON

Intelligence is the ability of the individual to adapt himself ade-quately to relatively new situations in life. PINTNER

Intelligence is an acquiring capacity. WOODROW

In addition to these definitions of intelligence, two theories of the nature of intelligence should be mentioned. The first is the " two-factor " theory of Spearman. According to this theory, innate intelligence depends upon a general common

[1] "Intelligence and Its Measurement: A Symposium," in *Journal of Educational Psychology,* Vol. XII, pages 123–147, 195–216; March and April, 1921.

factor which contributes to each specific performance of the individual. He conceives of this common factor as consisting of a general fund of mental energy, which may be focused upon a specific task. Spearman and others consider the common factor as a single trait that may be inherited from the parents in the same way that physical traits, such as the color of eye, are inherited. The second theory is the one advocated by Thorndike.[1] According to this theory, general intelligence consists of many innate capacities more or less closely related. These related capacities, he believes, may be grouped under three main types of intelligence: (1) abstract intelligence, which functions in connection with abstract ideas and linguistic skill; (2) mechanical intelligence, which functions in dealing with things; and (3) social intelligence, or the capacity to understand people and get along with them. The three types of intelligence just listed he believes to be positively related though not necessarily to a high degree. According to Spearman's theory, differences in the achievement of a pupil in different subjects, such as arithmetic, language, geography, etc., may be accounted for on the basis of differences in the opportunity offered him to become proficient in these subjects plus the differences in general intelligence and special abilities.

Evaluation and interpretation of concepts of intelligence. It is beyond the scope of this book to attempt a critical comparison and analysis of the concepts and theories of intelligence mentioned above. We may note that while the authorities quoted are not in perfect agreement in their statements, they nevertheless agree that intelligence is important in thinking and in adaptation to new situations. This is equivalent to saying that it is important in learning, which is one of the conditions that makes a knowledge of the intelli-

---

[1] E. L. Thorndike, "Intelligence and Its Uses," in *Harper's Magazine,* Vol. CXL, pages 227–235; January, 1920.

gence of a child very important to the teacher. As a practical matter, it is not necessary to know the exact nature of intelligence in order to measure it or to recognize the way in which it may function, any more than it is necessary to know the exact nature of electricity in order to measure and use it.

We have seen also that intelligence is largely a matter of endowment, as suggested especially in such studies as those of the Kallikak, Jukes, and Edwards families. The hereditary character of intelligence may be seen most clearly in such families as these, which deviate strikingly from the average or normal; but we should not assume that average intelligence is not also inherited. We should also be careful not to assume that individuals may be definitely classified into a few groups or levels, such as the feeble-minded, the normal, or the genius; this subject is discussed more fully in the following chapters. When the intelligence of large numbers is tested, there are no gaps or breaks. As is true in the case of other human traits, the distribution of intelligence is continuous.

It should be emphasized also that whether we accept the theory of general intelligence advanced by Spearman or by Thorndike, we should not expect an individual to perform evenly in all intellectual work. Specifically, if a pupil's general level of intelligence is high, we may expect high achievement in the various school subjects, though not necessarily equal attainment in all of them. Similarly, if a pupil's level of intelligence is low, we may expect a low rate of achievement in different subjects, though not an equally low rate in all subjects. If special traits and aptitudes are involved, we may look for even greater discrepancies in the performance of the individual.

Special traits and aptitudes. It is now pretty well established that an individual may inherit special traits and apti-

tudes, aside from general intelligence, that make it relatively easy for him to become proficient along certain specialized lines. For an individual to have such special traits and aptitudes may be as important to him and to society as for him to have a high level of general intelligence. Among such traits may be mentioned musical talent, mechanical aptitude, desirable traits of will and character, of temperament and emotion, positive moral qualities, good physical traits, etc. Such traits as these, together with general intelligence, may be combined in all sorts of ways so that literally no two people are exactly alike. It is clear, then, that even if we succeed in obtaining an exact measure of a person's general intelligence, important as that would be, we may even then be a long way from the true measure of his potentialities until we have obtained accurate measurements of other important traits and characteristics that he possesses.

Tests for measuring intelligence. Intelligence tests may conveniently be classified as individual or group tests. Individual tests are those administered to one person at a time; group tests may be given to a number of persons at the same time. Each type has advantages and disadvantages. In this chapter individual tests will be discussed, and in the following chapter group tests.

The Binet-Simon Tests. We have already referred to the pioneer work of Binet and Simon with intelligence tests. As already stated, their first scale (the 1905 scale) was developed in response to a definite need in the study of subnormals. The scale consisted of thirty separate tests designed to measure general intelligence. However, these tests were not standardized on an age basis. As a result of defects observed by the authors themselves, the scale was revised and standardized on an age basis in 1908. In other words, the authors attempted to group under each age the tests that best measured the intelligence of that age. This

arrangement made it possible to interpret a child's intelligence in terms of the average or normal child in his age group. This interpretation involves the use of age norms, which Terman[1] considers as one of the most important discoveries, from the practical point of view, in the history of psychology. Binet's final revision of the scale, extended to include tests for adults, appeared in 1911. Doubtless this would not have been Binet's final revision if his death had not occurred in the same year.

The Stanford Revision of the Binet-Simon Tests. Other psychologists were quick to see the possible practical uses of the Binet Tests, which were, therefore, soon introduced into America and into many other countries. Several American psychologists found that the tests needed further revision and modification for use in this country. Of these revisions, the Stanford Revision by Terman and his assistants, published in 1916, has been the one most widely used in the United States. We shall therefore describe it in some detail in order that the student may appreciate the nature and use of the scale and the careful and painstaking work underlying its construction.

Summary of the Stanford Revision. Using Binet's 1911 scale of fifty-four tests as a foundation, Terman added thirty-six new tests, twenty-seven of which were devised by him and nine by other psychologists. The ninety different tests were placed in age groups ranging from age three to " superior adult." Before discussing the method used to select and standardize the tests for each age, we shall list, for illustration, the tests finally adopted for ages three, eight, and twelve.[2]

[1] L. M. Terman, *The Measurement of Intelligence*, page 41. Houghton Mifflin Company, Boston; 1916.
[2] Arranged from the " Abbreviated Filing Record Card " for the Stanford Revision of the Binet-Simon Tests. Used by permission of Houghton Mifflin Company, publishers.

Year III

1. Points to parts of body: nose, eyes, mouth, hair
2. Names familiar objects: key, penny, closed knife, watch, pencil
3. Pictures, enumeration or better: (a) Dutch Home, (b) River Scene, (c) Post Office
4. Gives sex
5. Gives last name
6. Repeats 6 to 7 syllables: (a) I have a little dog. (b) The dog runs after the cat. (c) In summer the sun is hot.
   Alternative. Repeats 3 digits: 6–4–1; 3–5–2; 8–3–7

Year VIII

1. Ball and field   (Inferior plan or better)
2. Counts 20 to 1   (40 seconds, 1 error allowed)
3. Comprehension: What's the thing for you to do:
   (a) When you have broken something which belongs to someone else?
   (b) When you are on your way to school and notice that you are in danger of being tardy?
   (c) If a playmate hits you without meaning to do it?
4. Gives similarities: (a) wood and coal; (b) apple and peach; (c) iron and silver; (d) ship and automobile
5. Definitions superior to use: (a) balloon; (b) tiger; (c) football; (d) soldier
6. Vocabulary, 20 words
   Alternative 1. Names six coins
   Alternative 2. Writes from dictation: "See the little boy."

Year XII

1. Vocabulary, 40 words
2. Abstract words: pity, revenge, charity, envy, justice
3. Ball and field   (Superior plan)
4. Dissected sentences: (a) for the started an we country early at hour. (b) to ask paper my teacher correct I my.   (c) a defends dog good his bravely master.
5. Fables: (a) Hercules and Wagoner; (b) Maid and Eggs; (c) Fox and Crow; (d) Farmer and Stork; (e) Miller, Son, and Donkey
6. Repeats 5 digits backwards: 3–1–8–7–9; 6–9–4–8–2; 5–2–9–6–1
7. Pictures, interpretation: (a) Dutch Home, (b) River Scene, (c) Post Office, (d) Colonial Home
8. Gives similarities: (a) snake, cow, sparrow; (b) book, teacher, newspaper; (c) wool, cotton, leather; (d) knife blade, penny, piece of wire; (e) rose, potato, tree

How the revision was made. The work of making the Stanford Revision extended over a period of several years. An indication of the amount of work required and the care with which it was done may be seen from the following outline:

(1) About 2300 subjects, including 1700 normal children, 200 defectives and superior children, and more than 400 adults, were examined. When it is recalled that each subject was examined individually, from thirty minutes to an hour being required for each subject, it may readily be seen that this part of the work in itself was no small task.

(2) All the results for each test of the scale, obtained by all workers in all countries, were assembled to the extent that this was possible. The collection, classification, and analysis of this material again represent a great outlay in time and work.

(3) Determining the standards by testing the children in communities selected as typical of the country as a whole was another important step. It may readily be seen that, unless care was exercised, communities or schools not typical of the country as a whole might have been included, with the result that the test would not have been properly standardized. The use of communities including large proportions of foreigners, negroes, or other atypical groups would have yielded inaccurate standards.

(4) A verbatim record was obtained for nearly all the subjects examined, making it possible to rescore the tests according to any chosen standard if a change became necessary.

(5) Half a year was given to training the examiners and another half year to supervising their testing. This was done to insure uniformity in the results.

(6) Finally, all the records were scored by Terman himself in order to secure uniformity of procedure in this phase of the work.

In determining in which age group a given test should be placed, the aim was to secure a result that would cause the median mental and chronological ages of a large group of unselected children to coincide. In terms of individual testing, this would mean that the average child of any given age, say 9 years, should test at this same age mentally. It is clear that, in order to obtain this result, a great deal of experimenting with the test items was necessary. Terman states that after the first draft of the revision had been tried and had been found unsatisfactory " three successive revisions were necessary, involving three separate scorings of the data and as many tabulations of the mental ages before the desired degree of accuracy was secured."

The use of mental age and the intelligence quotient. As has been indicated, Terman followed Binet's age method of expressing scores. In addition to this, he made use of the " intelligence quotient," or IQ as it is usually abbreviated. The IQ is obtained by dividing the mental age (MA) of the subject by his chronological age (CA). This may be expressed as follows: $\dfrac{MA}{CA} = IQ$. Let us suppose, for example, that a child with the chronological age 8 years and 9 months has been examined and has earned a mental age of 9 years and 3 months. Converting these ages into months and substituting in the above formula, we obtain: $\frac{111}{105} = 1.05$. The pupil's intelligence quotient is then stated as 105.[1] Thus we see that the IQ expresses in a definite number the ratio between the subject's mental and chronological ages. It is obvious that an IQ of 100 expresses exact normality,

[1] In order to clear the figure of decimals, it is customary to multiply the obtained IQ by 100.

while IQ's above or below 100 indicate intelligence above or below normality.

Interpretation of results. On the basis of test results secured in testing many subjects of all levels of intelligence, Terman [1] suggests the following classification of intelligence quotients obtained in using the Stanford Revision:

IQ	CLASSIFICATION
140 or more	"Near" genius or genius
120–139	Very superior intelligence
110–119	Superior intelligence
90–109	Normal, or average, intelligence
80–89	Dullness, rarely classifiable as feeble-minded-ness
70–79	Borderline deficiency, sometimes classifiable as dullness, often as feeble-mindedness
Below 70	Definite feeble-mindedness

The significance of the various levels of intelligence listed above can best be appreciated when the examiner has tested many individuals at each level. It is helpful in this connection to make frequency tables showing the distribution of IQ's of the subjects tested, as in Table 30. The scores tabulated in this table were obtained by the writer [2] in testing 880 school children, most of them from Grades I and II.

In order to give a more definite notion of the characteristics of pupils of the various intelligence levels, it may be helpful to describe the pupils in terms of their achievements in school. As indicated in Table 30, approximately half of the pupils are of normal or average intelligence. In the elementary grades these are the pupils who make progress at a normal rate under normal school conditions. They do average school work as indicated by school marks or by standardized

[1] L. M. Terman, *The Measurement of Intelligence*, page 79. Houghton Mifflin Company, Boston; 1916. By permission of the publishers.

[2] I. N. Madsen, "Some Results with the Stanford Revision of the Binet-Simon Tests," in *School and Society*, Vol. XIX, pages 559–562; May 10, 1924.

educational tests. They are the pupils to whom the course of study seems best suited. They ordinarily complete the work of the elementary grades at about the age of fourteen. Many of these pupils go on into high school.

TABLE 30

Showing Distribution of Intelligence Quotients Obtained from the Stanford-Binet Tests in the Elementary Grades

IQ	Boys	Girls	Totals
140–149	1	1	2
130–139	8	7	15
120–129	43	28	71
110–119	79	74	153
100–109	127	123	250
90–99	98	105	203
80–89	56	53	109
70–79	30	26	56
60–69	6	10	16
50–59	1	1	2
40–49	1	2	3
Totals	450	430	880
Medians	103	101	102

Pupils of " superior " intelligence are not so numerous, about 17 per cent of elementary pupils falling in this classification. These are the pupils who usually have superior school marks and do the work rather easily. Many of them are able to do the work of the eight grades in seven years, if permitted to do so. They constitute a large proportion of high school pupils. The " very superior " are less numerous, according to our table about 10 per cent falling under this classification. These are the pupils who do the required school work with comparative ease and yet have superior school marks. They frequently are able to skip one or more grades and to complete the work of the elementary grades

at the age of twelve or perhaps less, when teachers and parents permit them to do so. The " near " genius or genius group is even smaller, our table indicating that only two pupils out of 880 attained this rank. If permitted, such children are usually able to do first-grade work at about the age of four and to enter high school at ten or eleven.

Of the pupils below average, we find about 12 per cent classified under the term " dullness." These are the pupils who usually need a considerable amount of special help from the teacher to keep them from failing. The school marks of this group are frequently below average, and often they require nine years or more to complete the work of the elementary grades. They consequently finish the work of these grades, if they finish it at all, at the more advanced age of 15 or 16. " Borderline " deficiency is indicated by even greater dullness, though it is not usually sufficient to be designated as feeble-mindedness. Pupils in this group are frequently required to repeat the work of a grade. They are consequently old for their grade and seldom remain in school till they reach the eighth grade. Feeble-mindedness is indicated by intelligence quotients below 70, and only 2 or 3 per cent of school children in a typical school community fall into this group. Feeble-mindedness below moron grade is not frequently found in school, the imbeciles and idiots being so obviously defective that they are rarely entered in school at all. Even the morons are seldom found in grades beyond the sixth or seventh.

The constancy of the intelligence quotient. If the intelligence quotient is to be a useful measure of intelligence, it should not fluctuate noticeably when the test is repeated. In other words, if a child obtains an IQ of 80 in one test and 120 in the next test, we cannot be certain whether the child is " dull normal " or " very superior " in intelligence. Terman and numerous other workers with the Stanford Revision

of the Binet Test have investigated this matter and find that IQ's obtained from this test do remain fairly constant if the test is properly given. This constancy is a very important feature, for it enables examiners to make predictions based on the rate of mental growth in the individuals tested. Thus if a child of 6 years is tested and found to have a mental age of 4 years and 3 months with a resulting IQ of 70, we can predict that at 12 his mental age will be approximately 8 years and 6 months. This makes it possible to plan more definitely for the educational needs of the child.

One method of obtaining evidence concerning the constancy of a test is to compute the reliability coefficient as explained in Chapter IV. Kelley[1] quotes the reliability coefficients computed by Terman for the Stanford Revision for three groups of subjects as follows:

TABLE 31

SHOWING RELIABILITY COEFFICIENTS COMPUTED BY TERMAN FOR THE STANFORD REVISION OF THE BINET TEST

RELIABILITY COEFFICIENT	CHRONOLOGICAL AGE	NUMBER OF CASES	STANDARD DEVIATION
.92	8.0–9.0	108	12.4 months
.93	12.0–13.0	57	18.46 "
.93	Adults	180	24.6 "

Dickson[2] quotes reliability coefficients from thirteen different investigations that, with one exception, range from .82 to .95. Lincoln[3] reports an investigation in which thirty six-

[1] T. L. Kelley, *Interpretation of Educational Measurements,* page 294. World Book Company, Yonkers-on-Hudson, New York; 1927.
[2] Virgil E. Dickson, *Mental Tests and the Classroom Teacher,* page 66. World Book Company, Yonkers-on-Hudson, New York; 1923.
[3] E. A. Lincoln, "The Reliability of the Stanford-Binet Scale and the Constancy of Intelligence Quotients," in *Journal of Educational Psychology,* Vol. XVIII, page 626; 1927.

year-old children were given two Stanford-Binet examinations on the same day. He found the reliability coefficient to be .95 for both mental age and intelligence quotient. He also found the probable error of a score to be 1.3 months for mental age and 1.8 points for intelligence quotient. The results reported in these investigations are typical of many similar studies.

Learning to use the Stanford Revision. Anyone who desires to become proficient in the use of the Stanford Revision must be willing to spend considerable time in study and practice. First, he should acquire understanding of the scale, either from the study of Terman's explanation of it in his *Measurement of Intelligence* or from some other equally good source. Next, he should familiarize himself with the Condensed Guide and the Test Material. The procedure outlined should be strictly followed. Any deviation is likely to vitiate the results obtained.

The following summary presents the main steps [1] in administering the Stanford Revision that must be followed if the scores obtained are to be compared with each other or with the norms:

1. Secure the attention and effort of the child.
2. Insure quiet and seclusion during the examination.
3. Establish rapport with the child.
4. Keep the child interested and encouraged.
5. Avoid fatigue.
6. Adhere to the formula in giving the tests.
7. Avoid coaxing.
8. Follow strictly the directions for scoring.

The amount of actual experience and training that is necessary after the student is prepared to begin examining

[1] The student should read Chapter VIII of Terman's *Measurement of Intelligence* in this connection.

will of course vary for different individuals. The writer has found that most students must examine from twenty-five to fifty children in order to become familiar with every part of the scale and in order to acquire reasonable skill in administering, scoring, and interpreting the results. Terman reports experiments showing that five weeks of study and practice suffice to give fair accuracy. Dickson,[1] in supervising the training of teachers for examiners in a large school system, also shows that a like amount of training usually gives reasonably satisfactory results.

Other American revisions of the Binet-Simon scale. The Stanford Revision has been described in some detail with the idea that it may be taken as representative of the Binet-Simon method of testing intelligence. Much of what has been said of this revision is also true of the other American revisions. Therefore we shall briefly describe only two others. They are the Herring and the Kuhlmann revisions.

The Herring Revision. The Herring revision differs from the Stanford in that it is a point scale. The scale is divided into five groups of tests labeled from A to E. They are so arranged that it is not necessary for a child taking the examination to attempt all parts of every group. This is done to obviate the need for spending time on the tests obviously too easy or too difficult for the child in question. The tests themselves are similar to those in the Stanford Revision, but not identical with them. The scale is easy to administer and score, being with respect to scoring somewhat more objective than the Stanford. The score of a child is determined by the total number of points earned and may be converted into a mental-age score from a table of norms. The intelligence quotient is computed just as in the Stanford Revision.

[1] Virgil E. Dickson, *Mental Tests and the Classroom Teacher*, pages 198–202. World Book Company, Yonkers-on-Hudson, New York; 1923.

Herring [1] reports coefficients of correlation of .98 computed from groups of children examined by both the Herring and the Stanford revisions. Wilner [2] has computed intercorrelations between the five parts of the Herring Revision and the Stanford Revision that range mostly between .93 and .98. The high intercorrelations between the different parts of the Herring Revision give evidence of its reliability, while the high correlation with the Stanford Revision is evidence of its validity. Concretely, these facts indicate that any given child will receive very similar ratings when tested by these two scales. This makes the Herring Revision very useful as a check on the Stanford when there is any reason for a retest. The converse is also true.

The Kuhlmann Revision. Kuhlmann's first revision was made in 1912. This scale he extended and modified in 1922.[3] In the latter revision he standardized and included tests for children ranging in age from three months to two years. Above the age of two, the scale consists of eight tests for each age group, making a total of 129 tests in the entire scale. The Kuhlmann scale employs the mental-age method of stating scores, as do the Stanford and Herring revisions, and like them makes possible the computation of intelligence quotients.

Non-language intelligence tests. The Binet-Simon scale and its revisions all arrive at the intelligence of an individual through his knowledge of language. Thus, in using any of the American revisions, it is assumed that the subjects being examined have had normal opportunity to learn the English

[1] John P. Herring, "Herring Revision of Binet-Simon Tests," in *Journal of Educational Psychology*, Vol. XV, pages 172–179; March, 1924.
[2] Charles F. Wilner, "A Comparative Study of the Stanford and the Herring Revisions of the Binet-Simon Tests," in *Journal of Educational Psychology*, Vol. XV, pages 520–529; November, 1924.
[3] F. Kuhlmann, *A Handbook of Mental Tests.* Warwick and York, Inc., Baltimore; 1922.

language. The scales therefore may not be applied in testing foreigners who have not yet had such opportunity, or in testing the deaf. A number of ingenious scales have been devised that eliminate or reduce materially the use of language. The "form board" test originally devised by Seguin, for use in his study of the feeble-minded, illustrates one method of eliminating the language factor. This test uses a rectangular board about fourteen by twenty inches with a number of openings of various designs and a corresponding number of blocks which fit into the openings. The blocks are arranged in a predetermined order and the subject is asked to fit them into the openings as rapidly as possible.

A more modern illustration of non-language scales is the Pintner-Paterson performance scale, which measures the subject's intelligence by means of motor responses rather than language responses. The verbal directions given by the examiner are also reduced to a minimum. On the basis of the points earned, a mental age may be obtained. There are fifteen tests of the general type of Tests 1, 4, 8, and 12, which may be described as follows : [1]

> 1. *Mare and Foal Board.* This is a picture board of a mare and foal with a number of cut-outs which the subject has to put in the correct places. It is very simple and resembles a child's game and serves as a very good introduction for children. Time and number of errors are recorded.
>
> . . . .
>
> 4. *Two-Figure Board.* Nine pieces to be placed in two spaces. Time and number of moves are recorded.
>
> . . . .
>
> 8. *Healy Puzzle A.* Five rectangular pieces are to be fitted into a rectangular frame. Time and moves are recorded.
>
> . . . .

[1] Description quoted from R. Pintner, *Intelligence Testing,* pages 122–124. Henry Holt & Co., New York; 1923. By permission of the publishers.

12. *Picture Completion Test.* Subject is required to select the appropriate block out of many possible blocks to complete the picture. Quality of performance scored.

Advantages and limitations of individual tests. Among the advantages of individual tests we may list the following: (1) They are especially adapted to the testing of primary pupils who have imperfectly learned group coöperation. (2) They enable the examiner to observe individual peculiarities during the examination that may influence the performance of the subject. For example, a child may have sensory defects that escape detection in a group examination. Extreme nervousness, shyness, fear, and other emotional conditions are also more easily discovered in an individual examination. (3) Chance disturbances, such as the breaking of a pencil, copying, etc., can more easily be controlled in an individual examination. (4) They allow the use of a wider variety of tests and therefore measure general ability with greater accuracy.

On the other hand individual examinations are limited by the fact that there is a dearth of qualified examiners. As we have already seen, considerable training is necessary before such individual examinations as the Stanford Revision can be used with any certainty of reliable results. Even when qualified examiners are available, the amount of time required for testing each pupil usually makes it impracticable to examine all the pupils in a school system or even all those in the primary grades. For this reason individual tests are generally limited to special or problem cases, while the examination of whole grades or schools must be accomplished through group tests.

<div align="center">EXERCISES</div>

1. List the advantages of stating intelligence in terms of mental age.

2. Compute the intelligence quotients for each of the six pupils

in the table below. The CA's and the MA's are given in years and months; for example, 10–6 means 10 years and 6 months.

Pupils	CA	MA	IQ
1	10–6	10–8	——
2	11–3	10–9	——
3	12–8	9–0	——
4	9–4	13–1	——
5	11–7	10–10	——
6	10–2	10–6	——

3. Into which level of intelligence, according to Terman's classification, would each of the above pupils fall?

4. Select one of the two revisions of the Binet scale given in the text, excluding the Stanford. Summarize and report to the class the account given by the author of his methods of revision. What advantages does each author claim for his own revision?

5. Summarize and report to the class the main features of a non-language intelligence test other than those described in this chapter.

References

BUCKINGHAM, B. R. *Research for Teachers*, Chapter III. Silver, Burdett & Co., New York; 1926.

COLVIN, S. S. "Principles Underlying the Construction and Use of Intelligence Tests." *Intelligence Tests and Their Use* (The Twenty-First Yearbook of the National Society for the Study of Education, Part I), pages 11–44. Public School Publishing Company, Bloomington, Illinois; 1922.

COX, CATHERINE, et al. *Genetic Studies of Genius*, Vol. II, "The Early Mental Traits of Three Hundred Geniuses." Stanford University Press, California; 1926.

DEARBORN, W. F. *Intelligence Tests*. Houghton Mifflin Company, Boston; 1928.

DICKSON, VIRGIL E. *Mental Tests and the Classroom Teacher*, Chapter III. World Book Company, Yonkers-on-Hudson, New York; 1923.

DOLL, EDGAR A. "New Thoughts on the Feeble-Minded." *Journal of Educational Research*, Vol. VIII (June, 1923), pages 31–48.

FREEMAN, F. N. *Mental Tests*, Chapters IV, XVII, and XVIII. Houghton Mifflin Company, Boston; 1926.

GODDARD, H. H. *Human Efficiency and Levels of Intelligence*. Princeton University Press, Princeton, New Jersey; 1920.

HERRING, JOHN P. "The Nature of Intelligence." *Journal of Educational Psychology*, Vol. XVI (November, 1925), pages 505–522.

HERRING, JOHN P. "The Reliability of the Stanford and the Herring Revision of the Binet-Simon Tests." *Journal of Educational Psychology,* Vol. XV (April, 1924), pages 217–223.

HINES, H. C. *Measuring Intelligence,* Chapters I and II. Houghton Mifflin Company, Boston; 1923.

HOLLINGWORTH, LETA S. *The Psychology of Subnormal Children.* The Macmillan Company, New York; 1920.

—— *Special Talents and Defects.* The Macmillan Company, New York; 1923.

KOHS, L. C. *Intelligence Measurement.* The Macmillan Company, New York; 1923.

KUHLMANN, F. "The Kuhlmann-Anderson Intelligence Tests Compared with Seven Others." *Journal of Applied Psychology,* Vol. XII (December, 1928), pages 545–594.

LAUTERBACH, C. E. "Studies in Twin Resemblances." *Genetics,* Vol. X (November, 1925), pages 525–568.

LINCOLN, E. A. *Beginnings in Educational Measurement,* Chapter V. J. B. Lippincott Company, Philadelphia; 1924.

—— "The Consistency of Intelligence Quotients (A Case Study)." *Journal of Educational Psychology,* Vol. XIII (November, 1922), pages 484–495.

MERRIMAN, C. "The Intellectual Resemblance of Twins." *Psychological Monographs,* Vol. XXXIII (1924).

OTIS, ARTHUR S. "Reliability of Binet Scale and Pedagogical Tests." *Journal of Educational Research,* Vol. IV (September, 1921), pages 121–142.

PETERSON, JOSEPH. *Early Conceptions and Tests of Intelligence,* Chapters VII, VIII, IX, X, and XI. World Book Company, Yonkers-on-Hudson, New York; 1925.

PINTNER, RUDOLF. *Intelligence Testing.* Henry Holt & Co., New York; 1923.

PRESSEY, S. L. and L. C. *Mental Abnormality and Deficiency.* The Macmillan Company, New York; 1927.

TERMAN, L. M. *The Intelligence of School Children.* Houghton Mifflin Company, Boston; 1919.

—— *The Measurement of Intelligence.* Houghton Mifflin Company, Boston; 1916.

—— "Mental Growth and the I.Q." *Journal of Educational Psychology,* Vol. XII (September and October, 1921), pages 325–341, 401–407.

—— et al. *Genetic Studies of Genius,* Vol. I, "Mental and Physical Traits of a Thousand Gifted Children." Stanford University Press, California; 1925.

—— "Nature and Nurture: Their Influence upon Intelligence and upon Achievement." *Journal of Educational Psychology,* Vol. XIX (September, 1928), pages 362–409.

TERMAN, L. M. *Nature and Nurture: Their Influence upon Achievement* (The Twenty-Seventh Yearbook of the National Society for the Study of Education, Part II). Public School Publishing Company, Bloomington, Illinois; 1928.

—— *Nature and Nurture: Their Influence upon Intelligence* (The Twenty-Seventh Yearbook of the National Society for the Study of Education, Part I). Public School Publishing Company, Bloomington, Illinois; 1928.

WINSHIP, ALBERT E. *Heredity : A History of Jukes-Edwards Families.* Journal of Education, Boston; 1925.

CHAPTER SIX

GROUP TESTS OF INTELLIGENCE

The development of group tests of intelligence. In order to overcome the limitations of individual tests of intelligence noted in the preceding chapter and to permit a wider use of intelligence tests, psychologists soon began to experiment with methods of group testing. The first group test of intelligence along modern lines was one by Arthur S. Otis, then a student of Terman's. This test was ready for publication in 1917 when the United States entered the World War. Soon after this time a committee of the American Psychological Association was working on a group test of intelligence for use in the army. Otis very generously gave the results of his research to the committee, which made good use of the material. Thus group testing on a large scale, as we now know it, was first carried on in the army. After the war the army tests were extensively used for high school and college students until tests specifically adapted for school purposes were devised. Since many of these were modeled after the army tests, it may be well to describe the latter briefly.

The Army Alpha test of intelligence. Two group tests were developed for use in the army — Army Alpha and Army Beta. The first of these was designed for literate and the second for illiterate and non-English-speaking recruits. We shall first discuss the former.

The material of Army Alpha [1] is arranged under eight subtests, each having definite directions and time limits. There

[1] The student should have access to samples of this and other tests discussed. As much as possible, the administering, scoring, and interpreting of tests should be demonstrated to the class. A list of the more commonly used tests in the elementary grades, together with the names of the publishers, will be found at the end of the chapter.

is a total of 212 items in the test. The following list gives the title, time limit, and number of items for each sub-test. The kind of material used in the various sub-tests, which is similar to the kind used later in school intelligence tests patterned after Army Alpha, is illustrated on pages 122–125.

TABLE 32

Showing Titles, Numbers of Items, and Time Limits for Sub-Tests in Army Alpha Test

Title	Number of Items	Time Limits
1. Following directions . . .	12	(Each item timed)
2. Arithmetical problems . .	20	5 minutes
3. Practical judgments . . .	16	1.5 "
4. Synonym-antonym . . .	40	1.5 "
5. Disarranged sentences . .	24	2 "
6. Number series	20	2 "
7. Analogies	40	3 "
8. Information	40	4 "

The entire test, including the directions to the subjects, can be given in forty-five to fifty minutes. The scoring can be done rapidly and objectively by means of scoring keys. A feature of the test is that it is composed of five forms[1] of approximately the same degree of difficulty. This makes it possible to retest a subject without repeating the same form if, for any reason, it should be desirable to administer the test to the subject again.

In general the aim of the authors of the test was to include material drawn from the common experience of the men tested, so that the responses to the questions and exercises would not be greatly influenced by differences in the environment, such as home conditions, schooling, etc. Considerable

[1] A test is said to have more than one form when there are two or more arrangements of the test items that are similar but not identical. Such forms are theoretically of about the same difficulty.

experimentation was required before the final forms were developed. The results obtained in testing more than a million and a half men in the army proved that the Army Alpha test was a very useful test of intelligence. Neither the authors nor the psychologists who have used the test assert that it is infallible. However, it was found to be a more accurate instrument for determining the intelligence of a soldier than any method then obtainable. Furthermore the level of intelligence could be determined immediately upon the soldier's induction into the army, and he could be assigned without delay to the branch of the service for which he was best fitted.

Interpretations of scores on the Army Alpha test. The Army Alpha test, instead of stating intelligence in terms of mental age and intelligence quotient, states it in terms of point scores. After a sufficient number of men had been examined so that fairly reliable standards or norms could be set up, the following interpretation or meaning was given to point scores of the magnitude indicated:

TABLE 33

SHOWING INTERPRETATIONS OF POINT SCORES MADE ON ARMY ALPHA TEST

POINT SCORE	LETTER RATING	INTERPRETATION	PERCENTAGE DISTRIBUTION IN A REGIMENT
135–212	A	Very superior	6.0
105–134	B	Superior	12.0
75–104	C +	High average	20.0
45–74	C	Average	28.0
25–44	C −	Low average	19.0
15–24	D	Inferior	13.0
0–14	D −	Very inferior	2.0

It will be noted that the distribution for the regiment indicated above is approximately that of a normal frequency

distribution. The distribution was about the same for the army as a whole, although there was considerable variation between regiments. It will be noted also that the average intelligence was indicated by a point score between 45 and 74. For the sake of convenience, letter ratings were used instead of the point scores.

While we cannot go into great detail here to show what uses were made of this test in the army, a quotation taken from the directions to the army examiners [1] will indicate how it was to be used, as well as how it was not to be used: " The mental tests are not intended to replace other methods of judging a man's value to the service. It would be a mistake to assume that they tell us infallibly what kind of soldier a man will make. They merely help to do this by measuring one important element in a soldier's equipment; namely, intelligence. They do not measure loyalty, bravery, power to command, or the emotional traits that make a man ' carry on.' However, in the long run, these qualities are far more likely to be found in men of superior intelligence than in men who are intellectually inferior. Intelligence is perhaps the most important single factor in military efficiency, apart from physical fitness." Further: " Commissioned officer material is found chiefly in the A and B groups, although of course not all high-score men have the other qualifications necessary for officers. Men below C+ should not be accepted as students in officers' training schools unless they possess exceptional power of leadership and ability to command."

Important by-products of army testing. While the purpose of the Army Alpha tests, as has been stated, was to facilitate the classification of men according to their usefulness to the service, the data accumulated in examining so many men was also found to have great scientific and practical value. In

[1] See C. S. Yoakum and R. M. Yerkes: *Army Mental Tests*, page 24. Henry Holt & Co., New York; 1920.

the first place, the testing of the soldiers, who came from all walks of life, gave a more definite and concrete index to adult intelligence than had previously been accessible. A second by-product was the discovery of differences in the degree of intelligence found in the various vocational groups. It is clear that such differences, if they are marked enough, are important in connection with educational and vocational guidance. A third significant by-product was the discovery of racial and national differences. The army tests made it possible to make comparisons based on large samplings so that the conclusions reached may be considered fairly reliable. It is true that there has been a good deal of opposition to these conclusions, particularly from individuals and from groups whose social and economic theories would need to be modified if these conclusions are accepted as valid. The most impartial and careful studies of the army data, however, agree that even when all possible allowance is made for differences in the environment, there still remains substantial evidence that there are racial and national differences in intelligence of varying degrees of importance.[1]

The Army Beta intelligence test. When intelligence testing in the army began, it was soon discovered that there were many men so illiterate that they could not be given the Army Alpha test; furthermore, many of the foreigners could neither read nor understand the English language well enough to be tested by this means. To meet the need, the Army Beta test was devised. It is so constructed that it does not require that the subject have the ability either to read or to understand the English language. The test can be given in about fifty minutes and can be scored by means of keys or stencils. Since a number of non-language school tests have been patterned after Army Beta, we shall describe each of the

[1] Clifford Kirkpatrick, *Intelligence and Immigration.* Williams & Wilkins Co., Baltimore; 1926.

seven sub-tests. The type of items that appear in each sub-test is illustrated in Figure 11. These items constituted the fore-exercises for demonstrating the test to the recruits.

Test 1 — Maze test. There are five mazes. The subject must draw a line by the shortest route from left to right without going into any blind alleys.

Test 2 — Cube analysis. There are a number of drawings representing cubes arranged in regular order. The subject must state the number of cubes in each pile.

Test 3 — XO series. There is a series of different arrangements of the letters *X* and *O*. At the end of each series there are blanks in which the subject must complete the series.

Test 4 — Digit symbol. This is a substitution test in which the subject must substitute symbols for letters.

Test 5 — Number checking. There are two double columns of numbers in which the subject must check the pairs of numbers that are alike.

Test 6 — Pictorial completion. There is a page of pictures in each of which something is left out. The subject must fill in the missing parts.

Test 7 — Geometrical construction. There are ten items. Each item is made up of geometrical figures, one of which is a square. With the square are other figures which, if properly put together, make a square. The subject is required to show by drawing in the square how the other figures might be placed within it.

Army Beta yields a numerical score that is translated into a letter grade in the same way as the scores in Army Alpha. The point scores and letter ratings are interpreted as shown in Table 34 on page 120.

The use of Army Alpha in schools and colleges. After the war the army tests of intelligence were released for use by
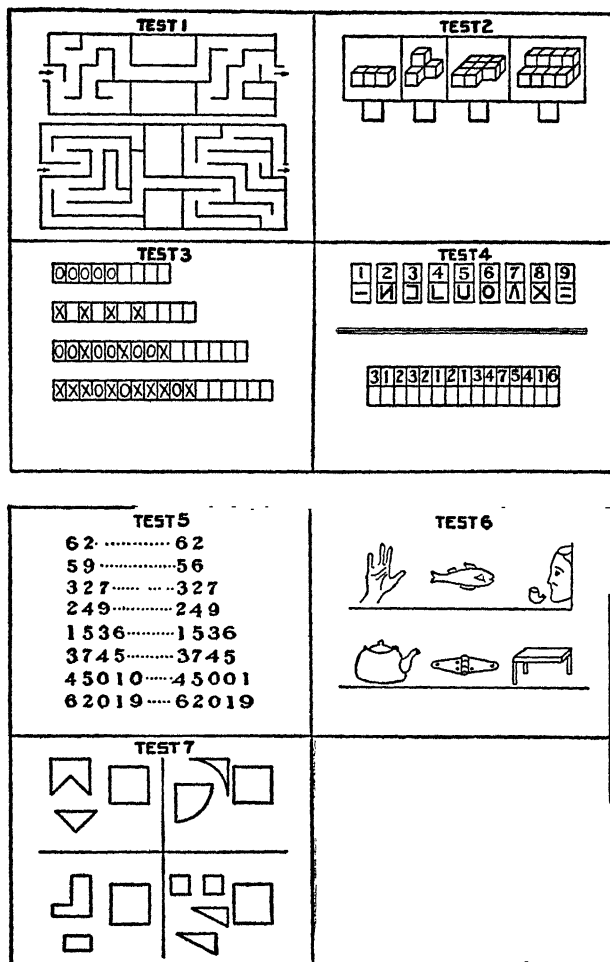
Fig. 11. Demonstration figures for Tests 1 to 7 of Beta as they appeared on Beta blackboard. (From C. M. Yoakum and R. M. Yerkes: *Army Mental Tests*, page 284. Used by permission of Henry Holt & Co., publishers.)

TABLE 34

SHOWING INTERPRETATIONS OF POINT SCORES MADE ON ARMY BETA TEST

POINT SCORE	LETTER RATING	INTERPRETATION
100–118	A	Very superior
90–99	B	Superior
80–89	C⁺	High average
65–79	C	Average
45–64	C⁻	Low average
20–44	D	Inferior
0–19	D⁻	Very inferior

civilian psychologists. Many schools and colleges took advantage of the opportunity, and thousands of students in high schools, colleges, and universities were given the Army Alpha test. Although this test was devised specifically for examining men in the army, it was found very serviceable in the types of schools mentioned. The educational uses of Army Alpha have, in general, been the same as those of other intelligence tests. These uses are discussed in detail in Chapter X. However, the early results obtained by use of Army Alpha in schools and colleges are of special value. They have given us the best norms for adult intelligence that are available with which to compare the intelligence of various school groups. The results are also valuable because they point the way to the construction and application of group tests more specifically adapted to school testing.

The development of group tests of intelligence for schools. After the war several psychologists who had previous experience in connection with the development and use of the army tests devised tests for use in the schools. Tests were soon available for testing pupils and students from kindergarten to university. At the present time probably not fewer than forty tests are in use for these purposes. At first the methods

used in devising and constructing intelligence tests followed the army methods rather closely, the attempt being chiefly to devise tests better adapted to the capacities of children in specific grades. In the space available here it is impossible and unnecessary to describe all the tests now in use. We shall therefore limit ourselves to the discussion of typical tests in elementary grades. The material included in such tests may be classified as linguistic and non-linguistic.

The linguistic material predominates in tests for the middle, upper, and high school grades. The assumption underlying the use of this material is that pupils of the same age or grade have had sufficiently similar life and school experience to insure that their intelligence will be represented with a fair amount of accuracy by scores earned in such tests. This assumption has proved well founded in most cases. But with a few pupils who vary noticeably from the normal in their living conditions and school experiences or who have been unusually limited in their opportunities to acquire a moderate knowledge of the English language the scores earned may not be indicative of their true intelligence. The material itself is usually arranged in a number of sub-tests, as in Army Alpha, in such a way that the difficulty of the items increases from the first to the last. Thus the youngest or the dullest pupils for whom the test is intended will be able to score on the first items, while even the oldest or brightest pupils will rarely be able to do all of the exercises. In this way it is possible to obtain measures for pupils of all levels of intelligence within the group for which the test is designed.

The non-linguistic material predominates in tests for kindergarten and primary grades. However, it is found to a limited extent in most linguistic tests and it is used entirely in a few tests for pupils of all grades. This kind of material is obviously useful in testing the intelligence of pupils who have not learned to read well enough to be tested by verbal

tests. To a great extent the same mental functions are tested as by the use of verbal material. The material used for testing pupils, whether it is verbal or non-verbal, is based on the same fundamental assumptions concerning the school experience of the pupils tested and concerning their life experience.

Organization of material in group tests. The following discussion deals with the most commonly used methods of organizing the material of a test under several sub-tests in which the items or exercises increase in difficulty from first to last.

1. *Directions.* In this type of test the subject is tested as to ability to comprehend and to execute directions. Either linguistic or non-linguistic material may be employed. For example, with linguistic material the subject may be asked to write numbers or letters in certain geometrical figures or he may be asked to cross out a certain letter in the words in which it occurs. With non-linguistic material the subject may be told to draw a line from one picture or figure to another by a direct or by a roundabout route, or he may be told to place dots or lines in a specified way with reference to pictures or figures, etc.

2. *Problems in arithmetic.* This type of material is found in many of the linguistic tests of intelligence. A pupil's score in this test is, of course, partly the result of schooling. The assumption is that when the schooling is approximately the same, differences in the number of problems correctly solved indicate differences in intelligence. The score is usually determined on the basis of the number of problems that are correct.

3. *Best answer.* This type appears in many of the modern intelligence tests. The usual form suggests three or more answers to a statement or question from which the subject selects the best, such as,

We go to school because —
1. Our teacher wants us to go.
2. We learn many useful things.
3. We have lots of fun.

4. *Opposites.* This type of test appears in several forms. It is used with both linguistic and non-linguistic material. A common method of arranging the material is:

If the two words mean the same, underline "same"; if they mean the opposite, underline "opposite":

quick..........slow	*same*	*opposite*
tall...........short	*same*	*opposite*
suspect........mistrust	*same*	*opposite*
allow..........permit	*same*	*opposite*

Another method gives a group of several words from which the subject selects one that is the opposite of a given word. For example:

Underline the word in the parentheses which is opposite in meaning to the first word:

reject.........(abuse, receive, accept, obtain)
often ....(seldom, allow, accept, obey)

In non-linguistic material the subject may be asked to make a mark to indicate the faster of two animals or machines or the larger of two objects, etc.

5. *Disarranged sentences.* In this type of test the words in a sentence are mixed up or disarranged and the subject is required to rearrange these words mentally to make complete sense. The response is usually made by underlining " true " or " false," as in the following:

If a statement is true, underline "true"; if it is false, underline "false":

all have wings horses	*true*	*false*
four have cats legs	*true*	*false*

6. *Completion tests.* This test occurs in a variety of forms. One form requires that the subject supply the omitted

words in a sentence, another requires that he complete a series of numbers, and a third asks the subject to mark missing parts of pictures. These three forms are illustrated in the following:

Write one word in each blank to complete the meaning:
    John has _____ to school.
    The snow _____ all night and we can _____ sleigh riding tomorrow.

Write the two numbers that should come next in each row:

2	4	6	8	10	12	__	__
7	1	6	1	5	1	__	__
16	17	15	18	14	19	__	__

Each of the pictures on this page has something missing, and you are to mark with your pencil the part which is left out.
(*This direction is followed by a number of pictures with missing parts.*)

7. *Analogies.* In this type the relation between a pair of words is given, and the subject is asked to point out another pair similarly related.

In each of the following lines the first two words are related in some way and you are to underline one of the words in the parentheses that is related in the same way to the third word:

    ice.. .cold   sun.. .(chair, man, rain, warm)
    dog....run   bird...(tree, sky, fly, boy)

8. *Information.* The purpose of this test is to obtain a measure of the subject's general information. The data may be drawn from either school or non-school experience. To be effective, it follows that there must not be too wide a discrepancy between opportunities the various subjects have had to acquire the information needed to respond to the test. The test may be based upon either linguistic or non-linguistic material. With the non-linguistic test, the subject may be asked to mark in some way common objects, or tell the time, etc. With lin-

guistic tests he is usually asked to indicate his information by underlining one of several words.

Underline the one word that best completes the sentence:

1. Paris is the capital of —
   Germany   France   Russia   Belgium

2. The carburetor is found in —
   a windlass   a typewriter   an automobile   a buggy

9. *Proverbs.* In one method the material is arranged by listing a number of proverbs, followed by a list of statements that explain their meanings. Each explanation is numbered, and the subject is asked to write the number of the explanation before the proverb that it explains.

10. *Memory.* Memory is of course involved in many of the preceding tests. However, it may be tested as a separate mental process by various devices. One method requires the subject to read a passage about which he must answer questions by underlining " yes," " no," or " didn't say." This test has also been used for primary pupils with non-linguistic material. The examiner holds a strip of paper showing four blocks. With a pointer he taps one or more of these while the pupil observes. The pupil is then required to mark on his test sheet the blocks which the examiner has tapped.

11. *Other kinds of material.* The foregoing descriptions of the types of material used in group intelligence tests do not include all the kinds of material now in use. They merely outline some of the more common types. New arrangements are constantly appearing that usually are modifications of those already in use. The descriptions do not do justice to the non-linguistic material which is difficult to describe or condense for reproduction. This type of material can be studied best from the tests themselves.

Organization for scoring.   The material of group tests may also be considered from the point of view of its organization for scoring.   Several methods have been referred to incidentally in our discussion of the types of material.   The various methods are discussed in detail in Chapter XI, in which this matter and others pertaining to the construction of objective tests will be considered.

The interpretation of group intelligence test scores.   There are several ways in which a pupil's score in a group test of intelligence may be represented and interpreted.   As in Army Alpha, many school tests state the scores in terms of the points earned.   This necessitates the use of age or grade norms as standards of comparison for individual scores.   Thus a pupil may be considered average for his age or grade if his score approximates the norm.   To facilitate interpretation of scores above or below the norm, percentiles are sometimes used, such as those for the Terman Group Test of Mental Ability shown on page 64.   From such tables one can readily determine just how a pupil ranks in relation to the other pupils of his grade.

One disadvantage of point scores is that they do not have the same value in different tests of intelligence.   That is, a score of 60 in one intelligence test is not necessarily the equivalent of a score of 60 in another.   When results of different tests are stated in terms of the point score, it is only by chance that the units of measurement are the same.   For a time it was thought that the conversion of point scores into mental ages and intelligence quotients, similar to those obtained from the Stanford Revision of the Binet Test, would result in equating the different tests.   For this purpose many intelligence tests provide tables from which any score can be converted into a mental age, which in turn can be changed into an intelligence quotient.   These are units of measurement that are easily derived and understood.   However, it soon

became evident that mental ages derived from various tests are not necessarily equivalents. Thus a mental age of 12 in one test may correspond to a mental age of 13 in another. It likewise follows that intelligence quotients derived from different tests are not necessarily equivalent.

The disagreement between mental ages from different tests comes partly from the fact that the variabilities from their respective means disagree. This causes the extreme scores of one test to vary less from their given mean than do the extreme scores of another test. Thus the mental ages of such tests are equivalent only at the mean. Another reason for the non-equivalence of mental ages in two different tests is the unreliability of such tests. Therefore errors in the scores of either or both of the tests are likely to cause differences in mental ages. A third reason for disagreement between mental ages is that the different tests have not been standardized on the same children. Thus if mental ages for one test are derived from a group of relatively bright children, and those for another from relatively dull children, it is obvious that the ages will not have the same meaning. Finally, different tests do not measure exactly the same kind of intelligence. This follows because different types of material are used in the various tests. However, there is no objection to the use of mental ages and intelligence quotients derived from a given test if we also name the test from which they were derived. For example, we might say that a pupil has a mental age of 12 years and 9 months according to the Terman Group Test of Mental Ability as compared with the mental age of 13 years and 6 months determined by some other group test.

The foregoing discussion of the use of mental ages and intelligence quotients applies only to pupils of those ages commonly found in the elementary grades. With students of the ages ordinarily found in high school and college the

mental age as a unit of measurement is of doubtful value. In the first place, innate mental growth stops during the teens. Terman estimates that this occurs at approximately 16, while other authorities give a lower or a higher age. In order to compute the intelligence quotient of a person beyond this age, one must divide his mental age by the exact age at which his mental growth ceased. Because the latter is a matter of conjecture, an individual might have an intelligence quotient that would be too high if the age of mental maturity had been underestimated or too low if it had been overestimated.

With superior adults the lack of adequate norms for comparison causes another difficulty. The mental age of a superior adult is significant in relation to that of the average person of his age. But if the age of maturity is taken at 16 for the average individual, we have no way of knowing the equivalent mental age of a superior person of this age. In other words, his point score in a group intelligence test cannot be given in terms of his true mental age because there are no age norms upon which to base the comparison. For these reasons it is better to interpret the intelligence of older pupils or adults in terms of the point score.

The validity and reliability of group tests of intelligence. As was explained in Chapter IV (pages 75–76), the validity of a test refers to the degree to which it measures what it claims to measure, while reliability refers to the accuracy with which it measures whatever it does measure. These are matters toward which test authors and test users are rightly directing more and more of their attention. The validity of a group test of intelligence is usually determined by correlating it against tests already established in the field, such as the Stanford Revision, or against other tests of recognized validity. We have seen in Chapter IV that the general reliability of an entire test can be determined from reliability

coefficients and related data, while that of a given score can be determined from its probable error. If these facts about a given test are obtained by a teacher, it is not probable that she will accept the results of testing uncritically. She will know, for example, that an obtained score of 70 with a probable error of 5 will indicate that the chances are even that the obtained score does not differ from the true score by more than ± 5. As already stated in Chapter IV, information concerning validity and reliability should be given in the manual of directions with other descriptive matter that accompanies a given test. If it is not found there, it may usually be found in references similar to those listed at the end of this chapter.

General rules for administering a group test.[1] In Chapter II the importance of adhering to the standardized procedure in the administration and scoring of standard tests was stressed. While the procedure varies somewhat for different tests and must be learned specifically for each test, there are certain common rules that it is useful to keep in mind. The author first stated these rules to aid teachers and administrative officers who had little or no training in the use of standard tests. The rules apply to intelligence and to achievement tests. They are arranged from answers given by the author to frequent questions from teachers and others. They are the rules and principles that are common to most tests now in use and are not intended as a substitute for directions accompanying the various tests but are intended to clarify and to supplement them.

1. If you have never before used the test you desire to give, study the directions carefully before giving the test. If possible, try the test on some other teacher or simply read

[1] These rules and the statement of "Reasons for Testing" in the following section are adapted from: I. N. Madsen, *A Teachers' Guide for the Use of Standard Tests*, pages 1–5. (Test bulletin published and copyrighted by the author; 1925.)

the procedure aloud in your room several times. It is important that you should be thoroughly familiar with the directions before giving the test to your class. Cultivate a pleasant but businesslike manner.

2. Pupils should look forward to taking a test without fear or nervousness. Most children like to take these tests if they are properly given. Never take the attitude: " Now I am going to get something on you." In getting ready for the test, have the pupils clear their desks and see that each has at least one well-sharpened pencil. It is well to have ready several extra pencils to distribute to pupils who break the pencil point. In case several grades are to take the same test, time is saved by moving the pupils to a large room. When this is done, care should be taken that the seats and desks are the proper size for the pupils that are to occupy them. Pupils should not be seated so closely that cheating or copying is encouraged. The tests should be given in a quiet room free from noise and disturbance.

3. Most tests have a definite time limit. This limit should be strictly adhered to and is particularly important in the short speed tests where the time varies from two to five minutes. In such tests a few seconds more or less may invalidate the results. If a stop watch is not available, a watch with a second hand should be used and the examiner should practice timing to the second. The exact time when the test was begun should be recorded as well as the time when the test was completed.

4. The directions for scoring should be thoroughly mastered. If you are in doubt about any point, consult some other examiner who has used the same test. The key which accompanies the test should always be followed rigidly, even if it seems arbitrary; otherwise your results cannot be compared with the standard scores.

5. When all of the tests have been scored, it is wise to check over all the arithmetical computations, such as subtraction and addition. As indicated in Chapter II, such errors are quite common when the examiners are inexperienced. It is wise also to have someone check over a few papers in order to determine whether the correct method of scoring was used.

Reasons for testing. In general a test should be selected for the purpose of solving a specific and definite problem. Often a test may contribute to the solution of several problems, and sometimes unexpected but valuable information is obtained as the by-product of a testing program. Unless there is a definite purpose, however, the testing is likely to result in little more than satisfaction of idle curiosity. Naturally problems of administrative officers and of teachers differ somewhat, though they overlap. Illustrations of both types of problems follow.

1. Problems primarily of interest to superintendents and principals
   - (a) The sectioning of classes into two or more sections according to ability. This may be done by the use of intelligence or educational tests or by a combination of both.
   - (b) The selection of pupils for special classes for exceptionally bright or exceptionally dull pupils, or the class arrangement of pupils with deficiencies in certain subjects.
   - (c) The determination of the efficiency of the school as a whole by comparison of obtained scores with standard scores, and with those made by other schools.
   - (d) The determination of whether the proper emphasis is given to all subjects.

    (*e*) The comparison of different methods of instruction or the comparison of new methods with old methods.

2. Problems primarily of interest to teachers

    (*a*) The determination of the efficiency of a class in the different subjects. Specifically, " Is my class up to standard in arithmetic, history, geography, reading? "

    (*b*) Similar determination of the efficiency of pupils in a class.

    (*c*) The determination of whether different subjects are given the right emphasis. Specifically, " Am I giving too much or too little time to arithmetic, spelling, geography, in comparison with other subjects? "

    (*d*) The diagnosis of defects of pupils in the various subjects. For example, " What phases of arithmetic need more attention? of history? of handwriting? "

    (*e*) The determination of whether a pupil is doing as well as can be expected. This requires the use of both intelligence and educational tests. A pupil who scores high in an intelligence test should score equally high in an educational test. A pupil who scores low in an intelligence test cannot reasonably be expected to score high in an educational test. In other words, a pupil's achievement is satisfactory if it is up to his level of intelligence.

EXERCISES

    1. Obtain several group tests of intelligence for the elementary grades. Classify each as to whether verbal or non-verbal material predominates.

    2. Estimate the relative proportions of verbal and non-verbal material used in each of the tests obtained.

3. Attempt to find methods of organizing the material in intelligence tests other than those described in the text.

4. Give a group test of intelligence to one or more grades. After scoring the tests, make frequency tables to show the degree of variability in each grade.

5. Compute the mental age for each pupil. Are the pupils properly classified according to their mental ages?

6. Give the same pupils a group test of intelligence different from that used in Exercise 4. Compute mental ages for this test. How do the mental ages in the two tests compare?

A List of Group Tests of Intelligence for the Elementary Grades

Army Group Intelligence Examination, Alpha. By a special committee of the National Research Council. For upper grades and high school. Has also been used for college students. Forms 5, 6, 7, 8, and 9, each $3.00 per 100 booklets; Manual of Directions, 75¢; Scoring Stencils, $1.25; Specimen Set, 80¢. Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kansas.

Cole-Vincent Group Intelligence Test for School Entrance. By L. V. Cole and Leona E. Vincent. For kindergarten and Grade I. Consists of non-verbal material. One form, $5.40 per 100; Scoring Key, 35¢; Set of Stencil Cards, 35¢. Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kansas.

Dearborn Group Tests of Intelligence, Revised Edition. By W. F. Dearborn. Series I for Grades I to III and Series II for Grades IV to IX. Series I is entirely non-verbal, and Series II is nearly so. There are two parts to each series. $1.00 per package of 25; Manual of Directions, 25¢; Set of Scoring Stencils, 25¢. J. B. Lippincott Company, Philadelphia.

Detroit First-Grade Intelligence Test. By Anna M. Engel. A standardized non-reading test for the first grade consisting entirely of pictorial material. Form A, $1.10 per package of 25, including Manual of Directions and Key; Specimen Set, 10¢. World Book Company, Yonkers-on-Hudson, New York.

Detroit Advanced First-Grade Intelligence Test. By Harry J. Baker. A non-reading test for children having completed the first three months of school, up to and including low second grade pupils. Form A, $1.10 per package of 25, including Manual of Directions and Key; Specimen Set, 10¢. World Book Company, Yonkers-on-Hudson, New York.

Goodenough Intelligence Test. By Florence L. Goodenough. A group test for kindergarten and Grades I, II, and III, based on characteristic differences in the spontaneous drawings of children. 60¢ per package of 25, including Key and Mental Age Equivalents of Scores; *Measure-*

ment of Intelligence by Drawing (the author's book containing directions for administering and scoring), $1.80. World Book Company, Yonkers-on-Hudson, New York.

Haggerty Intelligence Examination. By M. E. Haggerty. Group tests standardized by age and grade. Delta 1, for Grades I to III, consists of six sub-tests, five non-verbal and one verbal. Delta 2, for Grades III to IX, consists of six sub-tests that are modifications of the army examinations. Delta 1, $1.25 per package of 25, Key 10¢; Delta 2, $1.10 per package of 25, including Key; Manual of Directions, 25¢. World Book Company, Yonkers-on-Hudson, New York.

Illinois General Intelligence Test. For Grades III to VIII. Consists mostly of verbal material. Forms 1 and 2, each $2.00 per 100. Public School Publishing Company, Bloomington, Illinois.

Miller Mental Ability Test. By W. S. Miller. For use in Grades VII to XII and with college freshmen. Three sub-tests dealing with verbal material. Forms A and B, each 80¢ per package of 25; Manual of Directions, 15¢; Specimen Set, 25¢. World Book Company, Yonkers-on-Hudson, New York.

Multi-Mental Scale. By W. A. McCall. For Grade III and above. The 100 items of the tests are contained on a single sheet. Form 1, $1.00 per 100; Manual of Directions and Scoring Stencils, 15¢. Bureau of Publications, Teachers College, Columbia University, New York.

National Intelligence Tests. Prepared under the auspices of the National Research Council of M. E. Haggerty, L. M. Terman, E. L. Thorndike, G. M. Whipple, and R. M. Yerkes. For use in Grades III to VIII. The material and method are similar to those used in the army examinations. It includes Scales A and B with Forms 1, 2, and 3 of each. Each form, $1.25 per package of 25; Manual of Directions, 20¢; Specimen Set, 45¢. World Book Company, Yonkers-on-Hudson, New York.

Otis Group Intelligence Scale. By Arthur S. Otis. Primary Examination, non-verbal, for Grades I to IV. Forms A and B, each $1.10 per package of 25. Advanced Examination for Grades V to XII or any adult. Forms A and B, each $1.25 per package of 25; Manual of Directions, 30¢; Specimen Set, 50¢. World Book Company, Yonkers-on-Hudson, New York.

Otis Self-Administering Tests of Mental Ability (Intermediate Examination). By Arthur S. Otis. For use in Grades IV to IX. Forms A, B, and C, each 80¢ per package of 25, with Manual of Directions and Key; Specimen Set, 30¢. World Book Company, Yonkers-on-Hudson, New York.

Pintner-Cunningham Primary Mental Test. By Rudolf Pintner and Bess V. Cunningham. A group test of mental ability for kindergarten and first and second grades. The test consists entirely of pictures. Form A, $1.25 per package of 25, with Manual of Directions and Key; Specimen Set, 15¢. World Book Company, Yonkers-on-Hudson, New York.

Pressey Classification and Verifying Tests. By S. L. Pressey. The Primary Test may be used in Grades I and II, the Intermediate in Grades III to VI, and the Senior in Grades VII and VIII. Primary Test, $1.50 per 100; Intermediate and Senior Tests, $1.25 per 100. Public School Publishing Company, Bloomington, Illinois.

Terman Group Test of Mental Ability. By L. M. Terman. For Grades VII to XII or college freshmen. Consists of ten sub-tests based on material similar to the army examinations. Forms A and B, each $1.20 per package of 25, with Manual of Directions and Key; Specimen Set, 15¢. World Book Company, Yonkers-on-Hudson, New York.

References

BRIGHAM, C. C. *A Study in American Intelligence.* Princeton University Press, Princeton, New Jersey; 1923.

DICKSON, VIRGIL E. *Mental Tests and the Classroom Teacher,* Chapter III. World Book Company, Yonkers-on-Hudson, New York; 1923.

FREEMAN, FRANK N. *Mental Tests,* Chapters VI and VII. Houghton Mifflin Company, Boston; 1926.

GOODENOUGH, FLORENCE L. *Measurement of Intelligence by Drawings.* World Book Company, Yonkers-on-Hudson, New York; 1926.

GREGORY, C. A. *Fundamentals of Educational Measurement,* Chapter IV. D. Appleton & Co., New York; 1923.

GUILER, W. S. "The Predictive Value of Group Intelligence Tests." *Journal of Educational Research,* Vol. XVI (December, 1927), pages 365–374.

HAGGERTY, M. E. "Intelligence Examination, Delta 2." *Journal of Educational Psychology,* Vol. XIV (May, 1923), pages 257–276.

HENMON, V. A. C., and STREITZ, RUTH. "Comparative Study of Four Group Scales for Primary Grades." *Journal of Educational Research,* Vol. V (March, 1922), pages 185–194.

HERRING, JOHN P. "The Verification of Group Examinations." *Journal of Educational Psychology,* Vol. XV (December, 1924), pages 596–602.

HINES, H. C. *A Guide to Educational Measurement,* Chapters IX to XII. Houghton Mifflin Company, Boston; 1923.

HULL, CLARK L. *Aptitude Testing.* World Book Company, Yonkers-on-Hudson, New York; 1928.

Intelligence Tests and Their Use (The Twenty-First Yearbook of the National Society for the Study of Education, Parts I and II). Public School Publishing Company, Bloomington, Illinois; 1922.

KEFAUVER, GRAYSON N. "Need of Equating Intelligence Quotients Obtained from Group Tests." *Journal of Educational Research,* Vol. XIX (February, 1929), pages 92–101.

LINCOLN, E. A. "Studies of Validity of Dearborn General Intelligence Examination." *Journal of Educational Psychology,* Vol. XIX (May, 1928), pages 346–349.

MORRISON, J. CAYCE; CORNELL, W. B.; and CORNELL, ETHEL. "A Study of Intelligence Scales for Grades Two and Three." *Journal of Educational Research*, Vol. IX (January, 1924), pages 46–56.

McCALL, WILLIAM A., et al. "The Multi-Mental Scale." *Teachers College Record*, Vol. XXVII (October, 1925), pages 109–120.

—— "Construction of the Multi-Mental Scale." *Teachers College Record*, Vol. XXVII (January, 1926), pages 394–415.

MILLER, W. S. "The Variation and Significance of Intelligence Quotients Obtained from Group Tests." *Journal of Educational Psychology*, Vol. XV (September, 1924), pages 359–366.

OTIS, ARTHUR S. "An Absolute Point Scale for the Group Measurement of Intelligence." *Journal of Educational Psychology*, Vol. IX (May and June, 1918), pages 239–261, 333–348.

PINTNER, RUDOLF. *Intelligence Testing*, Chapter VI. Henry Holt & Co., New York; 1923.

—— "Results Obtained with the Non-Language Group Test." *Journal of Educational Psychology*, Vol. XV (November, 1924), pages 473–483.

—— "The Pintner-Cunningham Primary Test." *Journal of Educational Psychology*, Vol. XVIII (January, 1927), pages 52–58.

PRESSEY, L. W. "A Group Scale of Intelligence for Use in the First Three Grades." *Journal of Educational Research*, Vol. I (April, 1920), pages 285–294.

ROOT, W. T. "Correlations between Binet Tests and Group Tests." *Journal of Educational Psychology*, Vol. XIII (May, 1922), pages 286–292.

RUCH, G. M. "The Speed Factor in Mental Measurement." *Journal of Educational Research*, Vol. IX (January, 1924), pages 39–45.

—— and KOERTH, WILHELMINE. "Power versus Speed in Army Alpha." *Journal of Educational Psychology*, Vol. XIV (April, 1923), pages 193–208.

SYMONDS, PERCIVAL M. *Measurement in Secondary Education*, Chapter IV. The Macmillan Company, New York; 1927.

THORNDIKE, E. L. "The Improvement of Mental Measurements." *Journal of Educational Research*, Vol. XI (January, 1925), pages 1–11.

—— "The Measurement of Intelligence." *Psychological Review*, Vol. XXXI (May, 1924), pages 219–252.

VIELE, ADA B. "A Study of Four Primary Mental Tests." *Elementary School Journal*, Vol. XXV (May, 1925), pages 675–681.

WHIPPLE, G. M. "The National Intelligence Tests." *Journal of Educational Research*, Vol. IV (June, 1921), pages 16–31.

YOAKUM, C. S., and YERKES, R. M. *Army Mental Tests*. Henry Holt & Co., New York; 1920.

CHAPTER SEVEN

ACHIEVEMENT TESTS

As noted in Chapter II, standard achievement tests were developed in response to the need for more objective tests in measuring the results of teaching than those afforded by the ordinary methods of examining pupils. Such tests are now available for all the subjects included in the elementary grades. Their number is now so large that it would be impossible to describe and evaluate all of them in the space available. The discussion will therefore be confined to typical tests for the elementary grades. The general principles that underlie the construction of standard achievement tests have been outlined in Chapter II and have been discussed in Chapters V and VI in connection with intelligence tests. The most technical phases are outside the scope of this book.

Types of tests. Most tests for elementary grade subjects may be classified under three types according to their purpose or function: (1) *general survey tests* that attempt to measure proficiency or skill in a subject or in some phase of a subject, (2) *diagnostic tests* that attempt to discover the weaknesses and strengths of pupils, (3) *practice tests* that attempt to furnish practice in those phases of each subject in which the pupil has not acquired the necessary skill. As we shall see in the following pages, these three phases of testing have been worked out more completely for some subjects than for others.

I. ARITHMETIC [1]

As mentioned in Chapter I, the first standardized test in arithmetic was devised by Dr. C. W. Stone, who in connection

[1] Samples of the tests discussed in Chapters VII and VIII or others like them should be available for demonstration and examination. Students should have as much opportunity as possible to give and score the tests.

137

with his test made a study of the nature of ability in arithmetic. He came to the conclusion that instead of a single general ability, there were many relatively independent abilities. His conclusions in this matter are now generally accepted. It is obvious that they are of great importance in deciding not only what to teach and how to teach it but also what to measure. The point of view is now generally accepted that to teach arithmetic provision must be made for drill or practice in each specific phase of the subject where proficiency is desired. According to this view a pupil may be proficient in the addition, for example, of exercises such as $8 + 7$ without being proficient at the same time in the addition of exercises like $18 + 7$. Similarly, the fact that a pupil can multiply 7 times 9 does not guarantee that he can multiply 9 times 7. Since this holds true for all types of operations in arithmetic, it is important in teaching this subject to have tests for every phase of it so that the needs of the entire class as well as of the individual pupil may be discovered.

In the attempt to do justice to all phases of arithmetic, many standardized tests have been constructed. Indeed the number is so large that it will be impossible to discuss all of them in this book, or any of them in great detail. Nevertheless, a short explanation of the nature and function of typical tests will help us materially to acquire an understanding of tests in the field of arithmetic. For convenience, we may classify them under three main groups: (1) those that attempt to measure attainments or proficiency in the operations of arithmetic, (2) those that attempt to measure attainment or proficiency in the solution of problems in arithmetic, (3) those that provide drill or practice for pupils according to their special needs. The first two types may each be subdivided according to whether it is their function to make a survey of the general efficiency of a school or to diagnose the individual needs of a pupil. It should be noted, however,

that functions of different tests overlap considerably, so that some cannot be placed exclusively in any one of the above classifications.

1. *Tests Dealing with the Measurement of Proficiency in Connection with the Operations of Arithmetic* [1]

The Woody Arithmetic Scales,[2] Series A. This series of tests consists of four scales, one for each process — addition, subtraction, multiplication, and division. They are intended for use in Grades III to VIII. The examples in each scale vary in degree of difficulty from those that can be performed by pupils in the lower grades to those that will tax the capacity of pupils in the upper grades. This difference is illustrated by the first two examples as compared with the last two in the addition scale. These are:

(1)	(2)	(37)	(38)
2	2	$16\frac{1}{3}$	
3	4	$12\frac{1}{3}$	$25.09 + 100.4 + 25 + 98.28 + 19.3614$
		$21\frac{1}{2}$	
		$32\frac{3}{4}$	

The time limit for each grade is twenty minutes, which permits most pupils to attempt all the examples that they have mastered. These scales reveal the skill a pupil or a class has attained in working with abstract numbers. The scales may also be used to discover the deficiencies of a pupil or a class. This may be done by making a double-entry table for a class, recording the examples that each pupil missed. Steps may then be taken to rectify the conditions that are found. In this respect the Woody Scales are superior to the tests that

---

[1] The author, publisher, and other information concerning each test referred to in this and the following chapter are given in the bibliography at the end of Chapter VIII.

[2] These scales have been revised and are now published both in their original form by the Bureau of Publications, Teachers College, New York, and as the Woody Arithmetic Scales, Van Wagenen Revision, by the Public School Publishing Company, Bloomington, Illinois.

are limited to one type of example for each of the four operations. On the other hand, the Woody Scales are too brief to permit complete diagnosis.

The Woody-McCall Mixed Fundamentals. This scale is intended for use in Grades III to VIII and, as the name suggests, consists of a mixed arrangement of examples from the four fundamentals. As in the Woody Scales, the examples are arranged in order of their difficulty. There are four different forms of an approximately equal degree of difficulty so that the test may be repeated without danger of distorting a pupil's score because of his previous acquaintance with the scale. There are 34 to 35 examples in each form. Norms are available for each grade. In general, the scale may be used for the same purpose as the Woody Scales. As a matter of fact, it was constructed by selecting suitable examples from these scales.

The New Stanford Arithmetic Computation Test. This is Test 10 of the New Stanford Achievement Test and is part of the New Stanford Arithmetic Test. It is standardized for use in Grades II to IX inclusive. There are sixty items ranging in degree of difficulty from those suitable for pupils in Grade II to those suitable for pupils in Grade IX. The items were carefully selected, first on the basis of analyses of leading textbooks and tests in order to discover all possible types of examples, second in accordance with the judgment of four judges, and third on the basis of trial testing in a number of representative schools. Other valuable features of this test will be mentioned in Chapter VIII in the discussion of the New Stanford Achievement Test of which it is a part.

The Schorling-Clark-Potter Arithmetic Test. This test has been standardized after experimental work covering a period of several years. It has been standardized for Grades V to XII. The test consists of 100 examples divided into six sections, one each on the four operations with whole

numbers, fractions, decimals, and denominators; one on fractions, decimals, and per cents; and one on miscellaneous items. Two forms of the test are published, equal in difficulty as well as similar in arrangement and content. Grade norms by half years for the total score and for each part are available as well as a table for transmuting scores to arithmetic grades. The median reliability of the test is .85.

2. Tests Dealing with the Measurement of Proficiency in the Solution of Verbal Problems

The New Stanford Reasoning Test. This is Test 9 of the New Stanford Achievement Test and is part of the New Stanford Arithmetic Test. It is standardized for use in Grades II to IX and consists of forty items increasing in degree of difficulty from first to last. The authors state that " the guiding principles in the selection of problems for the Arithmetic Reasoning Test were that the problems should be worth while, that they should require real interpretative ability and not be made difficult through mere computation, and that they should be so clearly stated that the test would measure ability to think in quantitative terms; that is, arithmetic ability rather than chiefly a language function or a verbal intelligence." The important question of reliability, number of forms, type of norms, etc., are discussed in Chapter VIII in our consideration of the New Stanford Achievement Test as a complete unit.

The New Stone Reasoning Tests. The Stone Reasoning Test, as has been stated, was the pioneer standardized test in arithmetic. It has recently (1927) been revised and extended for use in Grades IV to IX inclusive. There are two approximately equivalent forms, each consisting of twenty-one problems arranged in order of their difficulty. The tests yield two scores, one for " correct answer " and one for " correct reasoning." The point scores obtained

may be converted into such derived scores as T-scores and age and grade scores.

The manual of directions, scoring key, and diagnostic record sheet provide for the administration and interpretation of the tests. They are not speed tests and are reasonable in the amount of time allowed a pupil. There are two tables in the manual of directions — one for " correct answer " and one for " correct reasoning " — that give reliability measures for each grade in which the tests may be used. These data include the number of cases on which the computations are based, the reliability coefficient, the probable error of r, the standard deviation, the index of reliability, the probable error of a score, the probable error of a score divided by σ, and the probable error of the estimated true score. These data provide the necessary information for the evaluation of the tests in terms of their reliability for groups and for individual pupils.

The Monroe Reasoning Tests. These tests have been used extensively, partly because they were among the first available and partly because the arrangement is convenient for teachers with limited training in the use of tests. There are now three forms of an approximately equal degree of difficulty for use in Grades IV to VIII. Each form is divided into three tests, one for Grades IV and V, one for Grades VI and VII, and one for Grade VIII. Each test consists of fifteen problems so arranged that a pupil can do all his work on the test paper. The problems are carefully selected from representative textbooks and are evaluated according to their difficulty. Each problem is rated for (1) " correct principle " and (2) " correct answer." A numerical value is given to each problem in proportion to the degree of difficulty for each of the two elements just mentioned; that is, a difficult problem receives a higher numerical rating than an easy one. The comparative values for each problem are accurately

determined according to the performance of the pupils previously tested in connection with the standardization of the tests. No credit is allowed for " correct answer," however, unless the problem is also solved satisfactorily so far as " correct principle " is concerned. It is also possible to obtain a rate score. This consists of the sum of the " principal values" of the problems that have been solved correctly in ten minutes. In practice, this score is often ignored. The time limit for the test is twenty-five minutes. Norms are provided for each grade.

The method used to determine the score of a pupil for " correct principle " sometimes allows the subjective element to creep in, because it is practically impossible to list all the possible correct methods of solving the problems. Monroe provides a scoring key for this purpose in which a number of acceptable solutions are listed. A pupil may have used other solutions that have not been listed in the key; the scorer is left to judge whether such solutions are acceptable. This also makes the scoring tediously slow if it is done conscientiously. On the other hand, the scoring of a pupil's paper for " correct principle " permits a more adequate diagnosis of his performance than mere scoring for correct answer.

The Stevenson Problem Analysis Test. It is the function of this test to determine whether pupils can read and analyze problems. To the extent that it accomplishes this purpose, it is diagnostic. The test can perhaps best be understood by an illustrative problem taken from it.

On May 5th Alice deposited $0.50 in the school bank; on the 10th she deposited $1.50; on the 15th she put in $0.50; and on the 20th she deposited $1.00. How much did she deposit all together during May?

(   )   *A.*   Which of the following facts are given in the problem?
        1. The different amounts deposited.
        2. The total amount deposited.
        3. The interest paid by the bank.
        4. The time when the money was due.

(   )   *B.*   Which of the following things are you asked to find out in the problem?
      1. The profit gained on the deposits.
      2. The number of times that she deposited money.
      3. The total amount deposited.
      4. The amount of each deposit.

(   )   *C.*   Which of the following is the most reasonable answer?

1.	2.	3.	4.
$22.00	$15.00	$1.00	$3.50

(   )   *D.*   Which process should be used in solving the problem?

1.	2.	3.	4.
Addition	Subtraction	Multiplication	Division

The procedure is explained to the pupils in connection with the trial problem above. After the pupils understand what they are to do, they are told to proceed in the same manner with the remaining tests. Ample time is allowed to finish the tests; that is, this is not a speed test. There are two forms, each consisting of two separate tests, one for Grades IV–VI and the other for Grades VII–IX. Standards are provided for each grade. The test calls attention to an important phase of arithmetic which has largely been overlooked by teachers until recently; namely, the importance of the technical vocabulary in solving problems in arithmetic.

The Compass Survey Tests. These tests were constructed by Knight, Greene, Ruch, and Studebaker. They are intended for use in general surveys of achievement in arithmetic and they supplement other tests by the same authors, which are described in the following pages. There are two forms of equal difficulty for use in Grades II to VIII. The test items suitable for use in Grades II, III, and IV are arranged in the Elementary Examinations, Forms A and B. The items suitable for use in Grades IV–VIII are arranged in the Advanced Examinations, Forms A and B. The following schedule of sub-tests will indicate the scope of the Elementary and Advanced Examinations.

TABLE 35

Showing Scope of the Compass Survey Tests

Part	Elementary Examinations (Grades 2, 3, 4)	Advanced Examinations (Grades 4–8)
1	Addition	Addition
2	Subtraction	Subtraction
3	Multiplication	Multiplication
4	Division	Division
5		Percentage
6		General Problems

It will be noted that the Elementary and the Advanced Examinations overlap. This overlapping is to provide for variations among courses of study in different schools. It will be noted also that the Advanced Examination attempts to survey the whole field of arithmetic. The Compass Survey Tests are designed to enable a teacher to take stock of her class as a whole from time to time so as to permit reorganization on the basis of achievement. It is intended also as a preliminary step to further detailed diagnosis of both class and individual weaknesses found in the various processes of arithmetic. To this end the authors of the test have selected the test items so that arithmetic examples of all types are included. This, of course, is necessary in order to test the pupils adequately. Complete directions for the administering, scoring, and interpreting of the results are given in a manual of directions.

The Compass Diagnostic Tests in Arithmetic. This comprises one of the most comprehensive series of all arithmetic tests. It consists of two forms, A and B, for use in Grades II to VIII. The general purpose is to enable the teacher to discover a pupil's weakness in a given process and also his weakness in any step of that process. For this purpose there are

twenty different tests.   The comprehensiveness of the series is indicated by the following list :

Test I.   Addition of Whole Numbers.   Grades 2–8
Test II.   Subtraction of Whole Numbers.   Grades 2–8
Test III.   Multiplication of Whole Numbers.   Grades 2–8
Test IV.   Division of Whole Numbers.   Grades 4–8
Test V.   Addition of Fractions and Mixed Numbers.   Grades 5–8
Test VI.   Subtraction of Fractions and Mixed Numbers. Grades 5–8
Test VII.   Multiplication of Fractions and Mixed Numbers. Grades 5–8
Test VIII.   Division of Fractions and Mixed Numbers.   Grades 5–8
Test IX.   Addition, Subtraction, and Multiplication of Decimals.   Grades 5–8
Test X.   Division of Decimals.   Grades 6–8
Test XI.   Addition and Subtraction of Denominate Numbers. Grades 6–8
Test XII.   Multiplication and Division of Denominate Numbers.   Grades 6–8
Test XIII.   Mensuration.   Grades 7–8
Test XIV.   The Basic Facts of Percentage.
Test XV.   Interest and Business Forms.
Test XVI.   Definitions, Rules, and Vocabulary of Arithmetic. Grades 4–8
Test XVII.   Problem Analysis, Elementary.   Grades 5–6
Test XVIII.   Problem Analysis, Advanced.   Grades 7–8
Test XIX.   General Problem Scale, Elementary.   Grades 5–6
Test XX.   General Problem Scale, Advanced.   Grades 7–8

Each of the twenty tests provides for specific analysis of the steps or processes involved.   For example, Test IV, which deals with the division of whole numbers, provides for diagnosis of the following factors :

Part 1.   The Vocabulary of Division
Part 2.   Fundamentals of Short Division
Part 3.   Short Division with Carrying

Part 4. Multiplication, Addition, and Subtraction Used in Division in Parts 2, 3, 5, 6, and 7
Part 5. Estimating the First Quotient Figure
Part 6. Fundamentals of Long Division: Checking
Part 7. Finding Errors in Long Division

Each of the other nineteen tests similarly provides for analysis and diagnosis. A manual of directions is provided that explains in detail the construction of the tests and gives directions with regard to administering, scoring, interpreting results, and remedial steps which should follow the testing programs. The tests were devised by the authors of the Compass Survey Tests.

3. *Practice Tests in Arithmetic*

The logical step, after survey and diagnostic tests in arithmetic have been given, is to follow them with practice tests or exercises for the phases in which pupils have shown deficiencies. A number of practice tests are now available for this purpose. We shall here discuss only a few typical practice tests as illustrations.

The Schorling-Clark-Potter Instructional Tests in Arithmetic. This series of tests is by the same authors as the Schorling-Clark-Potter Arithmetic Test. It consists of a booklet for each of Grades V to VIII. Each booklet consists of a set of inventory tests followed by several practice tests. An inventory test covers several operations. If the pupil fails to attain a set standard on the test (the standard being not more than one example wrong), he takes the practice tests that follow. If he does meet the standard on an inventory test, he skips to the next inventory test. There are three goals for each practice test for bright, average, and dull pupils. The pupil may, with the agreement of his teacher, decide on any one of the three standards. If he fails to meet his standard, he repeats the practice test until he does reach it.

Thus the pupil's drill work is directed in accord with his needs. The drill in arithmetic may be made highly individualized by the use of this material. It is easy to handle since there is a uniform time limit of four minutes for all the tests, inventory and practice, in the series.

In the back of each booklet there is a set of diagnostic tests. They are accompanied by a chart showing which practice tests should be drilled on for each example in the diagnostic test missed by the pupil.

There is also an individual record form on which the pupil can keep a record of his progress, and tabular forms for the class on which the teacher can keep a record of the class as a whole.

The Courtis Practice Tests. This series consists of forty-eight sets of exercises dealing with the difficulties found in the four fundamental operations with whole numbers. A feature of the tests is that they are self-diagnostic. The pupils begin by taking a test which selects the operations on which they need practice. Following the location of weaknesses, appropriate remedial drill is given. The exercises are printed on cards covered with transparent paper on which the pupil writes his answer. Provision is made for the pupil to keep his own score, thus arousing his active interest in his progress. The tests are standardized so that the time allotment for the exercises on each card is uniform for the whole set. A time-saving feature of the test is that a pupil may score his own performance by reversing the card and reading the answers on the back. The ease with which the tests may be administered makes it possible for a teacher to conduct the drill as individual instruction, each pupil doing the exercises that will increase his facility in a specified phase of arithmetic.

The Studebaker Practice Exercises in Arithmetic. These cards are somewhat similar in function to the Courtis Practice Tests. They are intended to furnish general drill in the funda-

mental processes in arithmetic. The drill is provided by fifty cards that present each of the four fundamental operations in a mixed order. The exercises become more difficult as the pupil progresses through the set. Since each pupil advances at his own rate, the pupils will not finish the exercises at the same time. In other words, each pupil spends just the amount of time needed to acquire the necessary skill for each operation. The cards are ingeniously constructed so that the pupil writes his answer through perforations in the card on a sheet of paper underneath the card. When he has finished, he turns the card over and compares his results with the answers on the reverse side. The pupil is required to keep a record of his own progress from day to day. The teacher keeps a class record, which enables her to tell at a glance the degree of progress made by each member of the class.

The Economy Remedial Exercise Cards. These cards are devised by the authors of the Compass Diagnostic Tests and the Compass Survey Tests. They are designed to remove specific disabilities in working with whole numbers that have been revealed in the diagnostic tests. For this purpose forty-one cards are available. These cards are correlated with the first four diagnostic tests that deal with whole numbers, listed on page 146. In this way the teacher can remedy mistakes at once when diagnosis has been made. The exercises are printed on cards so perforated that a pupil can insert a sheet of paper under the card and write the answer immediately below each exercise. The answers are printed on the back of the cards so that each pupil can score his own performance. This permits individual instruction, as with the Courtis Tests, and allows each pupil to progress at his own rate. The authors of the test are planning similar cards that will deal in the same way with the skills required in the other sixteen Compass Diagnostic Tests.

The relation of testing in arithmetic to textbooks and courses of study. Makers of standard tests in arithmetic are confronted with the need of carefully examining and selecting their content. This practice has had a wholesome influence in creating the same attitude toward the content of textbooks and of courses of study. The recognition of the close relationship between testing and teaching forms the second important influence that is exercised by testing on the construction of modern arithmetic textbooks. This is well illustrated by the fact that standardized self-testing drills and remedial self-help material are now included as essential steps in the mastery of arithmetic. It is illustrated also by the use of supplementary workbooks that provide drill on the essentials as well as afford extra practice on persisting difficulties. A third important influence has caused teachers to recognize more definitely the need of providing for individual differences among pupils. Consequently textbook writers have attempted to determine how frequently the various types of examples and problems should be presented to the pupils. They have also tried to analyze the difficulties pupils are likely to have and to design practice material that will enlist and retain interest and effort, etc. The three influences just described may be clearly seen in such textbooks as the *Modern-School Arithmetic,*[1] the *Standard Service Arithmetics,*[2] and the *Searchlight Arithmetics.*[3]

II. READING

The present trend in the teaching and testing of reading. The outcomes to be sought in the teaching of reading have

[1] J. R. Clark, A. S. Otis, and C. Hatton, *Modern-School Arithmetic.* World Book Company, Yonkers-on-Hudson, New York; 1929. Grades III to VIII inclusive.

[2] F. B. Knight, J. W. Studebaker, and G. M. Ruch, *Standard Service Arithmetics.* Scott, Foresman & Co., Chicago; 1927 and 1928. Grades III to VIII inclusive.

[3] B. R. Buckingham and W. J. Osburn, *Searchlight Arithmetics.* Ginn & Co., New York; 1927. Grades III to VIII inclusive.

never been entirely agreed upon by those responsible for the elementary grade curriculum. Until recently the emphasis has been placed almost entirely upon oral reading. Through this the teacher attempted to develop skill in pronunciation, enunciation, and expression. For the majority of teachers and pupils they were the outstanding conscious aims in reading. In its extreme form, placing emphasis upon the vocal phases of reading caused much time to be spent on elocution, and the reading textbooks were consequently equipped with many selections that lent themselves to this purpose. It is true that teachers believed that there was a close relation between fluency in oral reading and comprehension of what was read. But the idea of measuring comprehension by any method other than oral reading did not occur to them.

In the light of modern research it appears that the objectives listed above are not the only ones, nor are they the most important ones to be attained. Indeed it was found that some of them could be attained more economically by other methods than those used in teaching oral reading. Other important discoveries showed little connection between fluency in oral reading and ability to get the thought from the printed page; so it was concluded that oral reading is of comparatively little use outside the oral reading class. Thus attention was turned to silent reading, which teachers had seldom considered important. Research soon revealed tremendous differences existing among the various pupils of a grade in speed and comprehension in reading. Some pupils not only read several times as rapidly as others but also comprehended much more of what they read. It was realized also that silent reading is important as a tool in mastering many other subjects in school. Thus the pupil whose skill in silent reading was most highly developed had a tremendous advantage over other pupils in mastering the subjects in

which silent reading functioned. Attention was likewise called to the fact that most reading done in life outside of school is silent and not oral reading.

The shift of emphasis to silent reading. Considerations similar to those mentioned have resulted in the shift of emphasis to teaching silent reading. However, it soon became apparent that this is a very complex subject, probably requiring different techniques for teaching and testing its various phases. Thus in reading material of a factual nature where securing information is of primary importance, such phases as the rate of reading, comprehension, organization, and retention of material are the important skills to be developed. On the other hand, in reading poetry and literature, such phases as appreciation, enjoyment, expression, and comprehension are important, while the rate of reading is of relatively little importance.

Theoretically it is possible to construct standardized tests for measuring each of the phases of reading listed above. In practice, however, the construction of such tests is limited largely to the measurement of rate and of comprehension in the reading of factual material. In some tests rate and comprehension are measured separately, while in others they are combined in a single score. Rate of reading can, of course, be determined by the amount done in a given time. In measuring comprehension, the three procedures most commonly used require placing the material in one or more of the following forms: (1) a vocabulary test for measuring ability to understand or define words, (2) a sentence test for measuring the ability to comprehend sentences, (3) a paragraph reading test for measuring the ability to understand or comprehend paragraphs. Much ingenuity has been exercised in the arrangement of the material and in the form of organization of each of these three types. This will be illustrated, in part, in the discussion of typical tests in the following pages.

It will be found also that of the three types of tests each one has its own advantages and limitations. The abilities measured by them are all related to the ability to comprehend a passage of factual material, but not to the same extent or in the same way. That is, if a pupil shows poor ability to comprehend a passage, it may be because of his poor vocabulary, his inability to condense the main idea into a single sentence, or his inability to grasp the main thought of a paragraph. In our discussion of diagnostic procedures we shall find that there are many causes for poor comprehension other than those that may be revealed by any of these three types of tests.

In so brief a discussion of the nature and function of reading, it is impossible to do more than outline some of the outstanding problems. More detailed accounts may be found in books dealing with the social and psychological factors involved. For our purpose we can best proceed by discussing the different types of reading tests now in use.

The Monroe Silent Reading Tests. These tests are among the first and the most widely used of all reading tests. There are three forms, each consisting of two tests. Test I is to be used in Grades III, IV, and V, while Test II is to be used in Grades VI, VII, and VIII. The exercises consist of short paragraphs taken from school readers and children's books. They are arranged according to their difficulty. Following each paragraph is a question on its content which the pupil is required to answer. Only five minutes are allowed for the whole test. The directions for the test are read by both teacher and pupils. The rate of reading is determined by the number of words read in a minute, and the comprehension is determined by the number of questions answered correctly. The scoring is done by means of a scoring key. The nature of the test is illustrated by the sample paragraph appearing in the instructions to the pupils:

I am a little dark-skinned girl.   I wear a slip of brown buckskin and a pair of soft moccasins.   I live in a wigwam.   What kind of a girl do you think I am?

Chinese   French   Indian   African   Eskimo

The Monroe Silent Reading Tests have been recently revised.   The time has been shortened to four instead of five minutes.   Some ambiguities in the first edition have been corrected.   The author of the test does not claim that it measures all phases of silent reading, but that its purpose is to " measure the ability of pupils to read simple paragraphs for the purpose of answering specific questions."   This limitation, if kept in mind, will prevent many erroneous interpretations of pupils' reading scores.   While this test has now been superseded by others that are more reliable for measuring the reading attainment of individual pupils, it has had a very useful function in acquainting numerous teachers with the objective method of measuring silent reading.   It has also been valuable in setting up norms for both rate and comprehension in reading by the grades for which it is standardized.   Thus it helps teachers to secure more definite notions concerning the normal degree of proficiency for pupils in the various grades.

The Burgess Picture Supplement Scale.   This scale is designed for Grades III to VIII.   The fact that there are four forms makes it possible to give the test at relatively frequent intervals.   In every form there are twenty paragraphs that give the pupil specific directions to follow in connection with the pictures that supplement each paragraph.   Five minutes are allowed in which to do the work specified in as many of these paragraphs as possible.   The point score taken from the results can be converted by means of a table into a " derived score " on a scale ranging from 0 to 100.   This scale is so arranged for each grade that 50 is the standard score.   An illustrative paragraph follows:

1. Here is a picture of a girl's head. Take your pencil and quickly draw a circle around the picture to make a frame for it. Do not spend time trying to make a very good circle, but draw it quickly the first time; then go on and read what the next paragraph tells you to do.

Teachers and others who use reading tests have liked a number of features of the Burgess Scale, which has consequently had an extensive use. Among the limitations of the scale may be mentioned its brevity, a factor that makes it relatively unreliable as an indication of ability of individual pupils. It is printed on a single large sheet and is therefore rather awkward for pupils to handle. Nor is it entirely objective in its directions to pupils or its instructions for scoring. Instead of giving two scores, as the Monroe Test does, the Burgess Scale yields a single score based upon the number of paragraphs read correctly in the given amount of time. In spite of the limitations pointed out, and possibly others not specifically mentioned, the scale has been of great value in directing attention to the more objective methods of measuring silent reading.

The Thorndike-McCall Reading Scale. This scale is designed for Grades II to XII. There are ten forms. The exercises composing each form consist of paragraphs arranged in order of their increasing difficulty, each paragraph followed by one or more questions to be answered by the pupil. This is not a speed test; therefore pupils may reread the exercises in order to determine the correct answer. The point scores made by pupils may be converted into derived scores that aid in interpreting the pupil's reading achievement. One of these is the T-score,[1] which is derived from the standard deviation of the distribution of the scores obtained by all twelve-year-old pupils regardless of their grade. The unit of this derived score is .1 σ (one tenth of the standard devia-

[1] A more detailed explanation of the derivation of T-scores will be given in Chapter IX.

tion). Since the distribution of reading achievement indicated by this scale is assumed to lie between $-5.0\,\sigma$ and $+5.0\,\sigma$, the division into tenths gives a total of 100 units. The scale therefore ranges from 0 to 100 in terms of the T-score. The T-score may also be converted into a reading age for each pupil. This is done by determining the average T-score for pupils of each chronological age. The two derived scores mentioned may be obtained for any given point score by referring to a table in the manual of directions for using the scale. Thus if a pupil's reading age, as determined by this scale, is found to be twelve years and three months, he is said to have the same reading achievement as the average pupil of that chronological age.

The longer time (thirty minutes) that is allowed the pupil and the greater length of the test makes power rather than speed the factor of reading that is measured. From the point of view of the classroom teacher the scale is easy to administer, score, and interpret. The exercises are sufficiently similar to the matter that a pupil meets in his everyday silent reading to give a fair measure of his actual ability in this phase of reading. The fact that the scale can be used in a number of consecutive grades makes it possible to set up for the different grades norms that provide definite and concrete information concerning the normal differences which may be expected in the achievement from grade to grade. Unfortunately the scoring, not being entirely objective, is rather slow and tedious. Also the reliability of the scale is relatively low, decreasing its usefulness for individual testing.

The Haggerty Reading Examination, Sigma 3. There are two forms of this test, each for Grades VI to XII. Each form consists of three parts, or tests: a vocabulary test, a sentence test, and a paragraph reading test. The time allowed for these three tests is five, three, and twenty minutes respectively. The directions for administering, scoring, and

interpreting the test are given in a manual that also provides norms according to age and grade. With three tests, each one measuring a different phase of silent reading, the examination offers some provision for diagnosis. The items in each of the three tests are arranged in order of their increasing difficulty. Each test provides for objective scoring, for the pupil must underline the correct words in the vocabulary test; he also must underline " True " or " False," according to which word he thinks best answers each statement in the sentence reading test; finally he must underline the words or statements that best sum up the contents of the paragraph he has read in the paragraph reading test.

The New Stanford Reading Test. The first part of this test is a comprehension test consisting of three pages divided into paragraphs. In each paragraph are blanks that must be filled to complete the meaning of the paragraph. There is a total of eighty such blanks. The attempt has been made to hold the interest of the pupils by giving useful information in each paragraph. A total of twenty-five minutes is allowed for this section of the test. The time allotment is again sufficient to make the test one of reading power or comprehension rather than one of speed. The second part of the test is a vocabulary or word meaning test consisting of eighty multiple-choice questions in which the words have been carefully selected and arranged. A total of ten minutes is allowed for this part of the test. In the reading test, as in the other parts of the New Stanford Achievement Test, the authors have taken unusual care to provide for ease in interpreting the scores. Other valuable features of all parts of the test will be discussed in Chapter VIII.

The Haggerty Reading Examination, Sigma 1. This test is designed for use in Grades I, II, and III. It has two parts. The first of the two parts is made up of twenty-five exercises, each requiring the pupil to carry out a direction, such as

" Put a stem on the apple " or " Put a cross on the ball." The second part consists of twenty questions that are either true or false. The pupil answers each question by under- lining the word " Yes " if he thinks the question true, and " No " if he believes the question false. Test 1 allows twenty minutes, while Test 2 allows two minutes. The directions for administering, scoring, and interpreting Sigma 1 are given in the manual of directions. This manual also contains norms for both grade and age.

Gray's Oral Reading Paragraphs. The testing of oral reading is still largely on a subjective basis. However, Gray has attempted to put the measurement of oral reading on an objective basis. His test consists of twelve paragraphs arranged in order of their increasing difficulty. The test may be used in Grades I to VIII. It attempts to discover six types of errors: complete mispronunciations, partial mis- pronunciations, omissions, substitutions, insertions, repeti- tions. Only one pupil may be tested at a time, and during the administration of the test no other pupils should be present. The examiner records the errors made as the pupil reads. Norms are available for each grade. Only one form of the test has been designed. The fact that only one pupil can be tested at a time necessarily reduces the amount of testing of oral reading that can be done. No doubt the scarcity of standardized tests for oral reading may be attrib- uted to the fact that we are still far from a definite knowledge of the legitimate functions of oral reading, and from a similar knowledge of the amount and kind of oral reading that should be done. Some authorities, for example, have held that excessive oral reading is detrimental to the effective teaching of silent reading. It has even been suggested that perhaps silent reading should be taught before oral reading.

Diagnosis in connection with reading. As is the case with other subjects already discussed, standardized tests should

lead to a detailed diagnosis of a pupil's performance. At present the tests available are chiefly limited to the survey type, although some permit rough diagnosis. However, the literature on the psychology of teaching silent reading gives many concrete suggestions for diagnosing a pupil's performance. Among the more common causes for poor silent reading, we may list the following:

1. Poor vocabulary
2. Excessive vocalization and lip movement
3. Too many fixations of the eye in reading
4. Regressive eye movements
5. Short span of attention
6. Short unit of visual recognition
7. Defective intelligence
8. Defective vision
9. Wrong training in phonetics

The list above merely illustrates some of the many causes of poor silent reading that may be discovered by adequate diagnosis. Starch [1] has attempted to analyze the steps or processes involved in the complete act of reading, as follows:

1. Reception upon the retina of the stimuli from the printed page
2. The range of the field of distinct vision on the retina
3. The range of attention in apprehending visual stimuli
4. Movements of the eyes
5. Transmission of the visual impressions from the retina to the visual centers of the brain
6. Establishment or arousal of association processes whereby the incoming impulses are interpreted
7. Transmission of the impulses from the visual centers to the motor speech centers

[1] Daniel Starch, *Educational Psychology*, page 302. The Macmillan Company, New York; 1927. Used by permission of the publishers.

8. Transmission of motor impulses from the speech centers to the muscles of the vocal chords, tongue, lips, and related parts

9. Execution of the movements of the speech organs in speaking words

Starch points out that only the first six are involved in silent reading. It is obvious, however, that each step is rather complex and may be analyzed into sub-processes. The actual and specific causes that contribute to deficiency in silent reading, if discovered at all, may not be observed until the teacher has made a careful and painstaking study of each individual case.

The relation of testing to teaching of reading. With the increasing realization of the importance of silent reading and the improvement in the technique of measuring its important phases, great changes have been made in both content and method. For example, flash cards are now used from the first grade on in order to establish good habits in eye movements, satisfactory attention, comprehension, etc. A veritable revolution has taken place in textbooks for reading. Great attention is paid, for example, to such matters as vocabulary, size of print, length and regularity of lines, content of selections, self-testing, etc.

III. HANDWRITING

The Thorndike Handwriting Scale. As stated in Chapter I, the credit for devising the first handwriting scale belongs to Dr. E. L. Thorndike. The scale described here was constructed for use in Grades V to VIII. It was constructed on the basis of general merit of a large number of samples of handwriting as rated by twenty-three to fifty-five competent judges. The samples comprising the scale are so arranged that they increase in merit in equal steps from a quality of 4 to one of 18. The worst sample, 4, was artificially produced

and is described as having " zero " merit, while the best sample, 18, was taken from a copybook. The intermediate samples were taken from the actual handwriting of school children. For each of several of the steps in the scale there is but one sample, while for others there are several samples. The chief purpose of the scale is to reduce the errors made by teachers in grading handwriting for general merit. The Thorndike Scale was soon in widespread use and is still rather generally employed. The detailed construction and use of the scale can best be understood from Thorndike's account in his booklet on *Handwriting*.[1]

The Ayres Handwriting Scale. The next handwriting scale was devised by Dr. Leonard P. Ayres in 1912. This scale was standardized on the basis of legibility. The handwriting that could be read the most rapidly was assumed to be the most legible. A number of selected and competent judges were accurately timed while they read a number of samples of handwriting. The results obtained by the various judges were then averaged and the scale was constructed. The first edition contained three styles of handwriting; namely, slant, semi-slant, and vertical. There were eight samples of handwriting for each style. As the moderate slant style of handwriting came to predominate over the other styles, the need for a multi-slant scale decreased. The 1917 edition of the Ayres Scale, also known as the Gettysburg Edition, contains the generally accepted moderate slant style. This edition has been reprinted again and again and has probably been used more extensively than any other handwriting scale. We shall therefore discuss it somewhat fully as representative of scales measuring general achievement in handwriting.

Giving and scoring the Ayres test. Detailed directions accompany the scale and of course should be followed in using

[1] E. L. Thorndike, *Handwriting*. Teachers College, Columbia University, New York; 1912.

it. The rate score is obtained simply by counting the number of letters written per minute by each pupil. The legibility or quality of a pupil's writing is determined by comparing it with the various steps on the scale and assigning to it the value of the step that it most nearly resembles. It will be noted that as this requires a degree of subjective judgment, different raters will not always agree exactly on the value of a given paper or sample. The disagreement is much less, however, than that shown between judges who do not use a scale. The accuracy with which the scale is used increases with familiarity and practice. When inexperienced teachers use the scale, it is well to have two or more teachers rate independently the same set of handwriting papers and then take the average of the ratings for each pupil. If two or more teachers are not available, a teacher can increase her accuracy by rating each paper two or more times and taking the average of these ratings.

Norms and their use. The Ayres Scale sets up for Grades II to VIII inclusive norms for legibility and also for rate of handwriting. These norms are shown in Figure 12. Such norms are valuable to the teacher in a number of ways. They make it possible for her to determine the efficiency of her instruction with reference to the individual as well as to her class or grade. With them the supervisor or principal can judge more accurately the efficiency attained in teaching different classes or grades, and can compare each grade or school with other typical grades or schools. This is obviously very useful information.

Handwriting scales may be used to discover how well a pupil ought to write in order to meet ordinary life requirements outside the school. Thus Freeman [1] found by his

---

[1] F. N. Freeman, "Handwriting," in *Minimum Essentials in Elementary-School Subjects — Standards and Current Practices* (Fourteenth Yearbook of the National Society for the Study of Education, Part I), pages 61–77. Public School Publishing Company, Bloomington, Illinois; 1915.
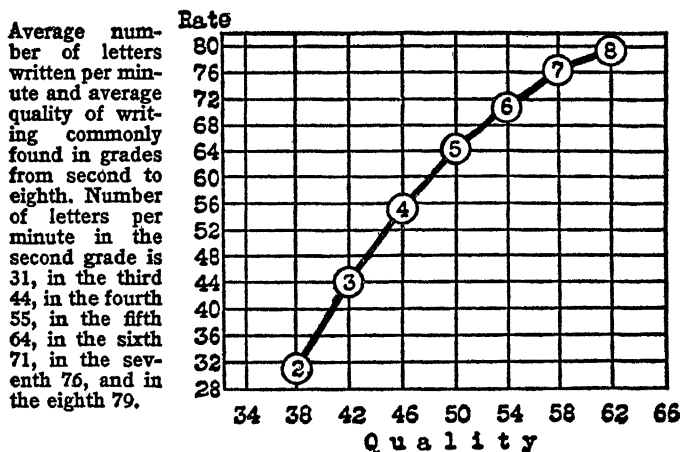
Average number of letters written per minute and average quality of writing commonly found in grades from second to eighth. Number of letters per minute in the second grade is 31, in the third 44, in the fourth 55, in the fifth 64, in the sixth 71, in the seventh 76, and in the eighth 79.

Rate

80 76 72 68 64 60 56 52 48 44 40 36 32 28

34 38 42 46 50 54 58 62 66

Q u a l i t y

Fɪɢ. 12. Showing graphically the Ayres norms for rate and quality in handwriting. (Reproduced from the *Ayres Handwriting Scale,* by permission of the Russell Sage Foundation, publishers.)

investigation among business men that they considered a quality of 60 on the Ayres Scale as satisfactory for ordinary business purposes. It will be noted with reference to Figure 12 that the average for the eighth grade is 62. In a more recent investigation Kirk[1] determined the handwriting proficiency found among people in a wide variety of vocations, using samples of both social correspondence and vocational handwriting. This he did by obtaining many specimens of handwriting from people in many different vocations. These specimens were then rated on the Ayres Scale to determine the average for each group. In this way he found that the average quality of handwriting in social correspondence ranged from 35.9 for physicians to 51.4 for housekeepers, with

[1] John G. Kirk, "Handwriting Survey to Determine Grade Standards," in *Journal of Educational Research,* Vol. XIII, pages 181–188, 259–272; March and April, 1926.

the average of 47 for all the groups compared. Similarly, he found that handwriting for vocational purposes ranged from an average of 44.4 for shipping clerks to 68.8 for elementary teachers, with the average of 54.4 for all of the groups. As Kirk points out, the averages obtained for the different vocational groups are not necessarily as high as the employers would like to have them, though they are not so low as to disqualify the employees concerned. Kirk next prepared a questionnaire in which he asked business executives what quality of handwriting they demanded. The average qualities demanded by thirty executives who answered the questionnaire were as follows :

Waybilling and tracing	75
Accounting	72
Post-office clerks	70
General clerical	65
Stock department	60
Factory	50
Retail-sales clerks	50
Hand-addressers	50
Telephone operators and supervisors	50
Messengers	42
Shipping and receiving clerks	35

The foregoing shows sufficiently how handwriting scales and norms may be used to establish more definite objectives or goals to strive for in the teaching of handwriting. It is obviously a waste of time to go far beyond social and vocational requirements, especially because in so doing time and effort would be taken from other school subjects. As a result of his study Kirk [1] sets up the standards for the various grades as indicated in Table 36.

While standards derived in the manner indicated above are not infallible objectives for pupils and teachers, they are far safer than standards set by subjective opinion.

[1] *Op. cit.*, page 271.

They should not, of course, be regarded as permanent; other investigations, changing requirements in the use of hand-writing outside the school, better methods of teaching hand-writing, or other forces may demand their modification.

TABLE 36

SHOWING GRADE STANDARDS IN HANDWRITING

	GRADES								
	I	II	III	IV	V	VI	VII	VIII	IX
Speed (letters per minute)	30	35	45	55	65	70	75	80	80
Quality (Ayres Scale)	35	40	45	50	55	60	65	70	75

Diagnostic scales and charts. The foregoing discussion has dealt primarily with the general achievement of pupils in handwriting. However, when a pupil's handwriting is discovered to be unsatisfactory, good teaching requires that the reason be determined as far as possible. For this purpose charts and scales have been devised that aid in the discovery of defects in writing. Thus Freeman has worked out a chart consisting of five scales to determine whether a pupil violates one or more of five essential traits or characteristics found in good handwriting. These scales are:

1. Uniformity of slant
2. Uniformity of alignment
3. Quality of line
4. Letter formation
5. Spacing

Each scale shows three qualities of excellence for the trait with which it deals — excellent, mediocre, and poor. Thus the teacher can take each pupil's handwriting and, by placing it alongside the scale, she can determine the quality from the point of view of slant, for instance. With a little

practice a pupil can be taught to analyze his own handwriting in this way. This procedure has the obvious advantage of enabling the teacher and the pupil to discover the weaknesses in handwriting that need special attention.

Many causes of illegibility in handwriting have remained unsuspected until recent times. Thus L. C. and S. L. Pressey [1] found approximately three thousand different illegibilities in the handwriting of children and adults. Among these there were 284 general characteristics that interfered with reading. Some examples of these characteristics are as follows: words crowded, writing too angular, rewriting, breaks between letters, loops too long, letters crowded, poor spelling, crowding at side of paper, etc. These writers also investigated to determine which letters of the alphabet caused the most trouble. Thus they found that the capital letters, with the exception of *I*, gave relatively little trouble. Among the small or lower-case letters, *r* yielded a total of 336 of the illegibilities, or 12 per cent of the total number. They found that the six small letters *r*, *n*, *e*, *a*, *d*, and *o* accounted for about one half of all the illegibilities. Table 37 shows the sixty-three malformations of letters and the frequency with which each error occurs.

In Table 37 the figures indicate the frequency for one thousand malformations. It is clear that charts made from such tables will be exceedingly valuable for diagnostic and practice purposes in handwriting. Lehman and Pressey [2] report an experiment in which they used a chart constructed from the foregoing material. A group of twenty-three 4B pupils were examined as to their illegibilities in handwriting.

[1] L. C. and S. L. Pressey, "Analyses of Three Thousand Illegibilities in the Handwriting of Children and Adults," in *Educational Research Bulletin*, Vol. VI, pages 270–273; September 28, 1927.

[2] Hilda Lehman and Luella C. Pressey, "The Effectiveness of Drill in Handwriting to Remove Specific Illegibilities," in *School and Society*, Vol. XXVII, pages 546–548; May 5, 1928.

TABLE 37

SHOWING SIXTY-THREE MALFORMATIONS OF LETTERS, WITH THE FREQUENCY OF OCCURRENCE OF EACH

a	like	u	30	n	like	v	11
a	"	o	24	n	"	s	5
a	"	ci	7	o	"	a	34
b	"	li	8	o	"	r	8
b	"	l	5	o	closed		6
b	"	k	4	o	like	u	5
b	"	f	4	r	"	i	59
c	"	e	18	r	too small		5
c	"	i	12	r	like	s	15
c	"	a	4	r	"	n	13
d	"	cl	41	r	"	u	8
d	"	I	5	r	"	e	4
d	"	a	4	s	indistinct		26
e	closed		52	s	like	r	10
e	too high		7	s	"	i	5
e	like	c	5	s	"	o	4
f	"	oj	5	t	"	l	19
f	"	b	5		cross above		11
g	"	y	13		no cross		7
h	"	li	19		cross right		6
h	"	p	9		cross left		5
h	"	b	7	u	like	oi	7
h	"	l	7	ur	"	w	4
i	"	c	10	v	"	n ,	6
	dot right		10	v	"	r	4
	dot left		10	w	"	u	8
	no dot		4	w	"	m	4
k	like	h	8	wr	"	ur	4
l	uncrossed	t	10	D	not closed		4
l	too short		8	I	like	cl	8
m	like	w	14	T	"	L	4
n	"	u	59				
				Miscellaneous letters			205
				Miscellaneous capitals			41

The teacher then met these pupils for remedial work for a half hour twice a week. The same teacher also taught a class of nineteen 3A pupils who took the handwriting drill given regu-

larly in the school. The conditions in the two groups were the same except that remedial help was given to the first group on their illegibilities in handwriting. Both groups were tested before and after the experiment in writing a dictation exercise and a composition, which were then rated on the Ayres Scale. The results of the experiment are shown in Table 38. In this table the group taught by the usual method is called the " control " group, while the one receiving remedial help is called the " experimental " group.

TABLE 38 *

MEDIAN STANDING OF A CLASS BEFORE AND AFTER TWELVE WEEKS OF DRILL IN LEGIBILITY, COMPARED WITH RESULTS FROM A CLASS NOT SO DRILLED

	DICTATION EXERCISE				COMPOSITION			
	SPEED		QUALITY		SPEED		QUALITY	
	Control	Experimental	Control	Experimental	Control	Experimental	Control	Experimental
Beginning	46	51	31	32	30	32	17	19
End	58	69	34	46	30	41	14	29
Gain	12	18	3	14	0	9	−3	10

*From Lehman and Pressey, *op. cit.*

Practice tests in handwriting. We have been shown how specific defects in handwriting may be discovered by the use of diagnostic tests and charts. Since different pupils will not always have the same difficulties, it follows that they will not need the same type of practice for their handwriting. Practice material should therefore be sufficiently flexible to give each pupil the kind of practice that he needs. One of the first men to devise a set of practice tests for handwriting was S. A. Courtis. The materials for the Courtis Practice Tests consist of a teacher's manual of directions, a pupil's daily lesson book together with a pupil's daily record card and graph, and a class record sheet. The first step in the use

of these practice tests is to give a preliminary or diagnostic test to discover the individual needs of the pupils. The procedure almost immediately changes to that of individual instruction because each pupil passes from one exercise to the next as soon as he has attained the standard agreed upon. Each pupil is shown how to direct his own progress by scoring and keeping records of his daily work.

Another example of practice tests in handwriting is found in the Leamer Diagnostic Practice Sentences. This system attempts to provide practice in writing letters and words that are most frequently used. The words and sentences used incorporate the words from the Ayres spelling list that are most often used in life outside the school. In one arrangement the set includes alphabet cards for each pupil, practice cards for sentences, a handwriting scale, a diagnostic chart of illegibilities in handwriting, and suggestions for follow-up work. One purpose of the practice material is to allow the individual to progress according to his individual ability.

The relation of handwriting tests to methods of teaching. As may readily be seen from the preceding discussion, the use of the diagnostic and practice test in handwriting makes it necessary to consider methods of teaching as well as systems of handwriting. A detailed consideration of these problems would be beyond the scope of this book. The relative merits of the various methods or systems of teaching handwriting must be determined partly on psychological and partly on social grounds. Psychology must tell us which methods are most effective in obtaining a desired result. Social considerations must tell us the results we should strive to obtain. However, we do not need complete or final solutions before we can improve present practices in handwriting. To attain this result, the tests we have discussed will be found very helpful.

CHAPTER EIGHT

ACHIEVEMENT TESTS (*Continued*)

IV. SPELLING

LIKE handwriting, spelling was one of the first subjects to be investigated. The investigations have dealt both with the testing of spelling proficiency and with the selection of words that should be taught. The work of Rice in this connection has been referred to in Chapter I. It has since been followed by a number of studies made with the purpose of selecting the words most commonly used in the English language. The work of the following investigators may be mentioned: (1) R. C. Eldridge, a New York business man, made a tabulation of words in four different newspapers. He found a total of 43,989 running words, which reduced to a total of 6002 different words. (2) W. F. Jones of the University of South Dakota obtained 75,000 themes from school children of four different states. He found a total of 15,000,000 running words written by 1050 pupils in Grades II to VIII inclusive. These reduced to a total of 4532 different words. The largest written vocabulary possessed by any one pupil was that of an eighth-grade girl who used 2812 different words. (3) L. P. Ayres of the Russell Sage Foundation tabulated words from 2000 letters dealing for the most part with business matters. In a total of 23,629 running words he found 2001 different words. (4) Daniel Starch tabulated words of forty authors found in eleven high-grade magazines. This yielded approximately 40,000 running words, or 5903 different words. Of these words, 3111 occurred only once each. That is, only 2792 occurred twice or more. (5) W. N. Anderson examined 3723 letters written by adults in various walks of life. He found a total of 361,184 running words and 9223 different words.

A number of similar studies have been made, but the examples given will suffice to illustrate the fact that the studies agree that there is a comparatively small number of words in common use. Word lists have been made by combining several such studies. Thus Professor E. L. Thorndike has compiled a *Teacher's Word Book* containing the 10,000 words most commonly used in the English language. This list shows the words most often met in reading. Similarly a list of the 10,000 words used most commonly in adult correspondence has been compiled by Ernest Horn of the University of Iowa. It is clear that lists of this type are extremely valuable in selecting words to teach. It is clear also that in making spelling tests and scales it is important to include the words most commonly used. It is obviously a waste of time to teach words that pupils will very rarely be called upon to spell outside the school.

The Ayres Spelling Scale. In constructing this scale, Ayres combined the words found in four word lists of the kind described above. Two lists were obtained by compiling words found in letters, one was based on words found in high-grade literature, and one on words found in newspapers. In this way Ayres obtained a total of 368,000 running words written by 2500 different individuals. From this list he selected the 1000 words that occurred most frequently. This list was divided into fifty sub-lists of 20 words each, to be spelled by pupils in various cities throughout the United States. Each list of 20 words was spelled by 70,000 pupils, making a total of 1,400,000 spellings or an average of 1400 spellings for each of the 1000 words. In this way the spelling difficulty of each word was determined as indicated by pupils in Grades II to VIII. The words were then arranged in twenty-six lists so that all the words in a given list were of about the same difficulty. The scale shows the percentage of pupils in each grade who should spell a given word correctly.

The Ayres Scale can conveniently be utilized in making spelling tests. The author suggests that a test should include twenty words, selected from a column of suitable difficulty for the grade to be tested. The numbers at the head of the column indicate the standard score for the grade or pupil tested. In order to make the scale more usable for the upper grades, B. R. Buckingham standardized 505 additional words and included them with the Ayres Scale. His edition is therefore called the Buckingham Extension of the Ayres Spelling Scale. But the words added were not selected on the basis of usage, and thus do not have the same significance as those of the original Ayres list.

The Iowa Spelling Scale. This scale is somewhat similar to the Ayres Scale. It was constructed by E. J. Ashbaugh from a vocabulary study by W. N. Anderson. The material comprised 3723 letters of personal correspondence by Iowa people. The scale differs from the Ayres Scale in that it includes nearly 3000 words (2977 to be exact) instead of the 1000 included in the Ayres Scale. Ashbaugh determined the spelling difficulty of all the words in the scale by having Iowa school children in Grades II to VIII spell them. Each word was spelled by two hundred or more children in each grade. The scale was originally published in three booklets, this arrangement making possible a more accurate placement of words. There is a separate scale for each grade in a more recent form of the scale. The larger number of words in the Iowa Scale makes it usable as a basic word list for teaching spelling, as well as for making up spelling tests for the various grades.

The Monroe Timed Sentence Spelling Test. The author of this test, W. S. Monroe, considers that the spelling of words in context, or in sentences, presents a more natural spelling situation than is presented by words dictated from a column; that is, it is more in accord with the requirements

for spelling in life outside the school. He presents evidence to show that the norms for words taken from the Ayres Scale are lower when the words are presented in sentences than when presented in column form. The words in the Monroe Test are taken from the Ayres Scale. There are three tests with 50 words in each, one for Grades III and IV, one for Grades V and VI, and one for Grades VII, VIII, and the high school. The words for the first of these tests are taken from Column M of the Ayres Scale, those for the second from Column Q, and those for the third from Columns S, T, and U. The time allowed for dictating the sentences was determined by the rate of speed in writing of more than six thousand Kansas school children. Specific directions for giving and scoring the tests, as well as norms for each grade, accompany the tests.

The New Stanford Dictation Test. This test is standardized for Grades II to IX inclusive. It makes use of the sentence method, considered superior by the authors because it approaches more nearly the requirements of daily life. It is given as a " dictation test " rather than a " spelling test," " for the sake of naturalness and in order to avoid the mental confusion which many pupils experience when they are conscious of the fact that they are being tested in spelling." Nearly all the words dictated count toward a pupil's score if spelled correctly, although there are a few connectives and fillers that do not count. The words are carefully selected on the basis of common usage and difficulty. They were taken from such word lists as the Ayres, Buckingham, Horn-Ashbaugh,[1] and " 7S." [2] Beyond the second grade there are three critical words in every sentence dictated. Pro-

[1] The Commonwealth List in *A Basic Writing Vocabulary* by Ernest Horn. (Published by University of Iowa, Iowa City.)
[2] From *Sixteen Spelling Scales*, prepared under the direction of Thomas H. Briggs and Truman S. Kelley. (Published by Teachers College, Columbia University, New York.)

vision is made so that pupils in upper grades do not need to write all the exercises designed for the lower grades.   The test is a part of the New Stanford Achievement Test and has the merits common to all the parts of that test battery.

Other spelling scales and tests.   There are other standardized spelling scales and tests which will not be discussed here because of limitations of space.   The main purpose of the preceding discussion has been to familiarize the student with the materials and the nature of such tests and scales.   The list at the end of the chapter includes the more commonly used spelling tests and scales together with those for other subjects.   A number of special considerations relative to the use of the various tests might be discussed.   For example : Should the test be presented in sentence or in column form ? At what rate should sentences be dictated ?   How should the material on which a grade is tested be selected ?   Concerning these considerations there is a fair amount of agreement among authorities on spelling tests and scales.   Some authors have made their decisions on such matters part of the test, so that when that particular test is used the teacher has a definite procedure to follow.   For the beginner in the use of standardized spelling tests, it is safe to say that the better-known tests give a far more accurate measure of achievement than could be obtained by the " homemade " spelling tests.

Diagnostic and remedial procedures.   As is true of handwriting, spelling scales and tests may be used to determine how a class or a pupil compares with the norms for the grade or the age concerned.   More specifically they may be used to diagnose weaknesses and to suggest remedies.   Remedial procedures may be taken directly in connection with teaching.   Thus the words that a pupil misspells are obviously words that need special attention.   In connection with these words the teacher may provide extra drill for the pupil and

may also direct him to keep a list of words that he frequently misspells, to be reviewed at intervals until they are mastered. This procedure should develop in a pupil a more self-reliant and critical attitude toward the results of his own efforts.

Whenever possible, the teacher should assist the pupil to discover why he misspells certain words. In this connection Dr. Jones lists one hundred words that are misspelled with the greatest frequency by pupils. This list is usually referred to as the " 100 spelling demons." The words are rather simple and common words, such as *which, their, there, separate,* etc. Such words may be called to the attention of the pupils so as to challenge their best efforts. Other causes for misspelling relate to the mispronunciation of words — as, for example, leaving the first *r* out of February as many persons do. Thus it becomes a natural matter to write " Febuary " in agreement with the mispronunciation. Whatever may be the causes for misspelling a word, good teaching aids the pupil to discover the specific type of error he is making as a starting point for his remedial work.

The relation of standard tests and scales in spelling to the course of study and to textbooks. The search for the most commonly used words in the construction of spelling tests and scales, noted in the foregoing discussion, has also caused curriculum makers and textbook writers to scrutinize with more care the words that pupils are required to learn. The demands of the modern curriculum are so great that only the most useful material can be taught. Thus the vocabulary studies referred to previously have resulted in a more careful selection of the words to be included in spelling textbooks. The number of words taught has been greatly reduced since the appearance of Rice's pioneer studies of spelling, referred to in Chapter I. A more careful placement of the word lists to be taught in the various grades is another result of the statistical studies made.

In this section tests will be considered for those phases of English usually designated as Language, Grammar, Composition, and Literature. The mere listing of these phases at once indicates the complexity of the subject. The construction of tests and scales to measure adequately the four phases mentioned presupposes, among other things, a decision upon their relative importance; for example, the relative importance of the functional versus the structural side of language — that is, language usage versus formal grammar. In general, we may say that increasing emphasis is being placed on language usage and less on formal grammar. As in other school subjects there must also be a close correlation between the subject matter taught and the material measured by the tests, if the tests are to give valid results.

The New Stanford Language Usage Test. This is a test of seventy-four items standardized for use in Grades IV to IX inclusive. The test is intended to measure two aspects of language usage, the choice of correct grammatical construction, and the choice of correct words for clear expression. Alternative choice of test items is the arrangement used. The time allotment is ten minutes. In selecting the test items, the results indicated in various investigations of children's language errors were considered. This test appears as Test 4 in the New Stanford Achievement Test (page 188).

The New Stanford Literature Test. This test consists of eighty items selected chiefly on the basis of a study of the investigations that have been made of the reading interests and practices of school children. Care was taken in selecting the items to appeal equally to the interests of both sexes, to have an equitable representation of items from American and foreign literature, to give due representation to authorship

and content and to the different classes of literature, such as travel, biography, etc. The multiple-choice organization of test items is used. The time allotment is ten minutes. The test appears as Test 5 in the New Stanford Achievement Test.

The Charters Diagnostic Language Tests. This test is the oldest of this type and has been used extensively. It includes four tests — Verbs, Pronouns, Miscellaneous A, and Miscellaneous B. For each test there are two forms. The tests may be used in Grades III to XII. Each test consists of forty exercises or sentences, most of which involve some language error. The pupil is required to correct the errors by making the necessary changes. He is allowed ample time to attempt all the exercises. The language errors that are included in the tests were selected from a study of errors made by school children in their written and oral work. The four tests combined make it possible to discover, in the case of a pupil or a class, which of the most common types of language errors are made. Remedial instruction may be based on the results obtained. Norms for each grade are available which make it possible to compare the efficiency of a pupil or of a class with the norms for the grade. The tests may be used either for diagnosis or survey or both.

With the exception of Miscellaneous B, these tests are issued in a form that permits the measurement of a knowledge of formal grammar. In this form the pupil is required to state the grammatical rule that is the basis for his correction of a sentence. In the original form the method of scoring was rather subjective; for many pupils stated the rule with ambiguity, making it difficult to draw the line between the acceptable and the unsatisfactory responses. The ambiguity of response also made the scoring slow and tedious. In the revised form of the tests the pupil is presented with a numbered list of rules from which he selects and records the number of the rule that justifies his correction. This modification

makes possible greater objectivity and speed in scoring.   The revised tests are designed for use in Grades VII to XII inclusive.

The Wilson Language Error Test.   The test consists of a short story, told in the form of a pupil's composition.   There are twenty-eight language errors, representative of the type frequently made by school children, which the pupil is required to correct.   There are six forms of the test — A, B, C, D, E, and F.   It is designed for Grades III to XII, may be used for diagnostic or for survey purposes, and is easily administered and scored.   Norms are available for all the grades.   Among the advantages claimed for the test by the author are included the following: (1) It is based upon comprehensive studies of the errors actually made by pupils and is therefore based upon the " right curricular material "; (2) it is put in the form of a pupil's composition, thus avoiding artificiality; (3) it is based upon sound psychological theory in that it helps the pupil to discover specific errors, so enabling him to work for a definite purpose; (4) it lends itself to economical teaching procedures by revealing the needs of each pupil to the teacher.

The Willing Scale for Measuring Written Composition. This scale is somewhat similar in construction to the Ayres Handwriting Scale.   It comprises eight compositions arranged in order of their increasing merit from a scale value of 20 to a scale value of 90.   The scale yields two scores termed " Story Value " and " Form Value."   The former is based on the quality and merit found in the story, and the latter is based on the number of mistakes found in spelling, punctuation, and syntax per hundred words.   The scale gives complete instructions for administering the test and for scoring the results.   The pupils are directed to write a story about an " exciting experience " they have had, or if they prefer, they may select instead a topic from a suggested list.

Twenty minutes are allowed for writing and five minutes are then given to finish, make corrections, and count the number of words written. The pupils' performances are scored by comparison with the scale. The scale may be used in Grades IV to VIII, and norms for each grade are provided.

New York English Survey Test — Literature Information.[1] This is part of the New York English Survey Tests for the measurement of achievement in language usage, sentence structure, grammar, and literature information. Since examples of the first three types of tests have been given already, we shall limit our discussion to the fourth test, Literature Information, designed for use in Grades VII and VIII. The test may be illustrated by one of the thirty-six exercises it includes. The first exercise is as follows:

1. *The Man without a Country* was written by —
       ☐ George Washington
       ☐ Edward Everett Hale
       ☐ Aaron Burr
       ☐ Philip Nolan

The pupils are directed to " Place a check in the square opposite the group of words that completes the sentence correctly." Five minutes are allowed for the test. The test provides for objective scoring and of course measures only one phase of literature — information. However, it illustrates the possibility of measurement in this complex field.

The relation between the testing and the teaching of English. As has been pointed out, testing in English as in other subjects, to be of any practical value, must be closely correlated with the information that is taught. As in other subjects, extensive research has been done in the field of English in order to determine the material that should be

---

[1] By J. S. Orleans, E. L. Cornell, W. W. Coxe, and E. B. Richards. Public School Publishing Company, Bloomington, Illinois; 1925.

included.　Reference has already been made to Charters'
study of language errors in the discussion of his scale for
measuring achievement in language (page 177).　To illustrate
types of studies made in this connection, we may quote from
Starch [1] a tabulation of language errors made by pupils in
various grades.

TABLE 39

CLASSIFICATION OF LANGUAGE ERRORS MADE BY PUPILS
IN VARIOUS GRADES

TYPES OF ERRORS	GRADES						ALL GRADES
	3	4	5	6	7	8	
1. Verbs	44.2	60.0	55.4	54.9	43.3	48.2	49.9
2. Pronouns	15.9	14.0	6.7	7.7	12.3	18.8	13.5
3. Negatives	11.5	7.1	20.2	7.3	15.2	14.0	11.6
4. Syntactical redundance . .	8.0	6.6	11.2	12.6	16.5	9.6	9.7
5. Mispronunciation	14.7	7.8	2.2	4.9		1.7	8.0
6. Prepositions	3.4	3.2	1.8	5.6	4.1	2.6	3.5
7. Adjectives and adverbs . .	2.0	0.6	2.2	6.6	8.2	4.8	3.3
8. Ambiguous expressions . .		0.2					0.2

Table 39 shows the per cent of different types of errors
occurring in each grade included in the study.　Thus in
Grade III, 44.2 per cent of the errors made were in the use
of verbs, 15.9 per cent were in the use of pronouns, etc.　This
is a summary table and does not show the frequency of
specific errors, although such tables have also been prepared.
They show, for example, how often such expressions as "ain't
got," "haven't got," etc., are used.　It is obvious that
studies of this kind will enable curriculum makers to organize
courses of study that will more nearly answer the require-
ments of pupils and that will also be an aid to textbook
writers.

[1] Daniel Starch, *Educational Psychology,* page 426.　The Macmillan
Company, New York; 1927.　Used by permission of the publishers.

VI. GEOGRAPHY AND HISTORY

The problem of measurement in these subjects is fundamentally the same as in the ones already considered. However, the question of the essential or fundamental requirements of the courses of study in these subjects is still a matter of considerable controversy. Thus it is obviously difficult for the test maker to be sure that his test meets one of the basic requirements of a good test; namely, that it is based upon the most important and useful phases of the subject in which the measurement is attempted. Nevertheless, much progress has been made and tests are available that, if properly used, will aid the teacher very materially.

The Courtis Supervisory Tests in Geography. The chief function of this series is to test the pupil's knowledge of certain facts in geography. Outline maps are included in the tests for the purpose of testing knowledge of locations. There are two tests, one concerned with the United States and the other with world geography. There are two forms for each test. The time limit for giving the test is five minutes. The procedures for giving and scoring are specific and easily understood. Such a test is limited in what it measures, and a pupil's score should not necessarily be taken as representative of his general knowledge of geography. The scores made by pupils and by classes in this test must be interpreted in relation to the emphasis that the teacher has given to each phase of the subject.

The Buckingham-Stevenson Place Geography Tests. There are two tests, one for the United States and one for the world in general, and there are three forms for each test. A rather unique feature of the test is that it is dictated to the pupils without the use of printed test forms or sheets. There are eighty questions for each test. The pupil merely takes a sheet of paper and writes down in a column numbers from

1 to 80. As the examiner reads a question, the pupil indicates his answer opposite the number corresponding to the question. The pupil's responses are scored by a key that insures the objectivity of the result. This test, as the name indicates, measures knowledge of locational geography. The tests may be used in Grades IV to VIII, and norms are available for each of these grades.

Some writers object to the use of such tests, and also to those of the Courtis type, because they emphasize a relatively narrow phase of geography and tend to encourage the mere memorization of facts. According to these writers the important phases of geography are related to the social problems that arise in connection with interrelations of nations and races. The objection does not seem to be especially well taken. It is obvious that a knowledge of facts is necessary for adequate thinking. Probably this knowledge should not be acquired by merely memorizing facts, as was the practice in an earlier day, but by studying the social problems that arise in the study of geography. However, the teacher must determine from time to time the progress made by the pupils in the subject as a whole and also in its various phases. Otherwise there can be no adequate diagnosis. Thus it may be that a pupil's reasoning in geography is faulty because it is retarded by inadequate factual knowledge. The fault lies not so much with the tool as with the use that is made of that tool.

The Gregory-Spencer Geography Tests. This test is designed for use in Grades VI, VII, and VIII, and has three forms. Each form contains eight sub-tests with a total of 111 test items. The sub-tests may be described by the following outline:

Test 1. Trade routes and products carried over them
Test 2. Miscellaneous exercises

Test 3. Causal geography pertaining to the United States
Test 4. Causal geography pertaining to the world in general
Test 5. The location of twenty-four cities of the world
Test 6. Descriptive phrases pertaining to the cities in Test 5
Tests 7 and 8. Location and description of countries of the world

The whole test may be completed by a pupil in thirty to fifty minutes. There is no definite time limit imposed on the pupil; he is allowed all the time that he can reasonably use. Specific directions for giving and scoring the test are provided, and the scoring is made objective by the use of keys. The authors of the tests have tried to make them representative of the subject matter taught in the grades for which the tests are intended.

The New Stanford Geography Test. This is Test 7 of the New Stanford Achievement Test. It consists of eighty multiple-choice items. The items were selected from analyses of textbooks and tests in geography. The test items cover a wide range of information, and care has been taken to insure that the facts of greatest social importance are represented. There is a time allotment of ten minutes for the test. It may be used in Grades IV to VIII inclusive.

The New Stanford History and Civics Test. The general make-up of this test is similar to that of the New Stanford Geography Test. The number of test items and their arrangement is the same. The items were largely selected in accordance with an analysis of eight leading social science textbooks. Ten minutes are allotted to this test. It may be used in Grades IV to IX inclusive. The scoring, as in the preceding test, can be accomplished rapidly and accurately by means of objective scoring keys.

The Gregory Tests in American History. These tests are constructed somewhat in the same manner as the Gregory-Spencer Geography Test. There are two forms of the test for Grade VII, two for Grade VIII, and two for the high school grades. Provision is made for objective scoring, and standards are available for the different grades. The nature of the test may be illustrated by listing the five sub-tests for Grade VIII, Form A. Test 1 deals with miscellaneous facts and dates in American history; Test 2, with the period of national growth from 1789 to 1830; Test 3, with the period of sectional dispute and the Civil War; Test 4, with the period of reconstruction and national development; and Test 5, with the period from 1900 to 1922. The tests have been criticized on the ground that they place too much emphasis on a knowledge of relatively unimportant facts. It is asserted also that the emphasis on facts in tests in this subject tends to encourage a method of teaching and of studying history that is now becoming obsolete. Since similar objections are urged against other factual history and social science tests, we can best consider the merits of these objections after the descriptions of the various tests.

The Hahn History Scale. This scale is arranged in much the same way as the Ayres Spelling Scale. It consists of a number of rather well-selected questions in American history which are grouped in columns of approximately the same degree of difficulty. Thus a teacher who desires to give a test in this subject can select questions for the grade to be tested from a column of suitable difficulty. At the head of each column is given the per cent that each grade should have correct. These are norms with which a teacher compares the results for her class. The scale is intended for use in Grades VII and VIII.

The relation between teaching and testing in the social sciences. History, geography, and civics are often referred

to as " content " subjects to differentiate them from arithmetic, spelling, reading, etc., which are referred to as " tool " subjects. The problem of testing in the social sciences is considered more difficult because the aims and outcomes that are desired seem more subtle. In the teaching of United States history, for example, it is frequently stated that the chief purpose is to give training in weighing problems, in order that the pupil may acquire a more " genuine " and intelligent patriotism and may become a more intelligent voter. Other important values to be obtained by a study of this subject are said to be ability to think historically, to judge and to evaluate historical facts and evidence, to deal with causal relations, and to appreciate the character of men. These values are regarded as more important than an acquisition of historical information. Insistence on the exact knowledge of historical dates, names of men, and other detailed facts is considered wrong in developing the main outcomes mentioned. It is further argued that an attempt to hold pupils responsible for knowing the detailed facts required in standardized tests will tend to distract both teachers and pupils from the more important aims and to focus their attention too closely upon the mastery or memorization of isolated facts.

The matter is further complicated because historians and interpreters of history do not agree on the actual meaning of patriotism or on its exact function. Nor do they agree on the interpretation and evaluation of historical events, or on the judgment of the character and motives of men, etc. These disagreements are real obstacles to the preparation of textbooks and courses of study. Until there is more agreement on what the content of United States history should be, and until the most important results expected from its study are decided upon, both testing and teaching will be very unsatisfactory. Unless we work toward fairly definite objectives

and have adequate methods of measuring our progress in their direction, our pupils are likely to be Topsies and " just grow," so far as attaining desirable outcomes from the study of history is concerned.

However, teaching cannot be separated completely from testing. The rejection of standardized tests would merely send us back to teachers' examinations, which, as we have noted, are extremely unreliable. If we also reject this type of testing, we are thrown back upon the still more unreliable general impression as a measure for the progress of pupils. In addition, if we are going to test with the purpose of discovering whether or not a pupil is making progress, and also with the purpose of modifying our procedure according to individual or group needs, the testing must be done parallel with the teaching or immediately following it. It may be very well to say that the chief aim in the teaching of history is to produce intelligent voting and genuine patriotism. But these functions are chiefly exercised after leaving school, and if our teaching has not brought about their proper development, it is then too late for the teacher to do anything about it. Her opportunity for testing and for taking remedial steps is limited to the classroom.

From the practical standpoint, if we teach history at all, measurement is reduced to an analysis of the content of the subject and of the relative emphasis given to each phase as it is actually presented to the pupils. On the basis of this analysis, the construction of tests may continue on the general principles used in constructing the tests for other subjects. Any other attitude would mean that we were willing, ostrich-like, to keep our heads in the sand and to know nothing about what our pupils attain from their study of the subject.

It is possible that the objection to the teaching of facts in history is not entirely sound, or that it has been misinterpreted by the classroom teacher. The psychological objec-

tion to the teaching and memorization of facts applies as much to the methods by which facts are taught as to the type of facts taught. Thus it was formerly the practice to teach and learn by rote the names of Presidents, lists of historical events, and dates, even though there was little requirement for a knowledge of these facts outside of school. In the tool subject of arithmetic an analogous situation existed when the multiplication table was memorized by rote. The method was objected to because the position of a given multiplication fact in a series was overemphasized, often making it necessary to recall the whole series before that required fact could be recalled. When it was discovered that the method was wrong, the facts of multiplication were not rejected but instead were taught in the relationships in which they were most likely to present themselves outside of school. In the modern school children often learn the facts of multiplication before they see a multiplication table. Drill is not eliminated but is used with better results because it is more in harmony with the laws of learning. As a pupil progresses in his study of arithmetic, the facts become associated with an increased number of problems. Such progress depends, to a large extent, upon a mastery of these and other facts usually designated as the fundamentals in arithmetic.

Is it not likely that facts in history may be similarly related to the solution of problems and to effective thinking in this subject? For example, in judging the character of Washington, many facts must be associated with him, including the conditions, customs, and habits of the people among whom he lived, his early life and training, the names and types of people with whom he associated, specific examples of how he dealt with problems and men, wars and battles in which he fought and the way in which he acquitted himself in them, activities and controversies in which he engaged during his presidency. In the same manner the evaluation of historical

evidence, the analysis of causal relations, and the like, require the use of many detailed facts. Generalizations in history are rich in meaning only when drawn from a wealth of facts with which they are closely associated.

It would seem, then, that facts *must* be learned in history. The problem is to correlate them with the main aims of the subject. There is no reason why teachers should not observe this need closely in their use of factual tests. If there is danger that teachers and pupils will exaggerate the importance of detailed facts by using such tests, there is equal possibility that they will obtain from some of the more extreme criticisms of factual testing and teaching now current the distorted idea that facts are unimportant. In this way the teaching of history might easily become little but froth and foam.

VII. OTHER ELEMENTARY GRADE TESTS

The New Stanford Achievement Test.[1] This test first appeared in 1923 in two forms, A and B. Each form, a booklet of twenty pages, contained nine separate tests grouped under six different subjects. Revision of the test was begun in 1925 and the new test, issued in 1929, consists of a twenty-four-page booklet of ten separate tests organized under seven different subjects. There are five forms — V, W, X, Y, and Z.[2] In the revision the outstanding change is the extension or improvement of each of the separate tests. The original form combined history and literature in one test, but in the new form they are separated and civics is included with history. The nature-study and science test of the original test has been replaced by tests in geography and physiology and

---

[1] Class discussion of this test will gain greatly in concreteness and clearness if the students have an opportunity to observe the actual giving of the test and scoring of the results.

[2] Only two forms, V and W, were published at first, the plan being to make the others available as needed. Forms V, W, and X are now available.

hygiene. However, the norms for the new test are comparable with those of the original one.

Three of the tests — reading, arithmetic, and dictation — may be used in Grades II to IX inclusive. All the others may be used in Grades IV to IX inclusive. There are two examination booklets for each form, the Primary for Grades II and III, and the Advanced for Grades IV to IX. Several of the subject tests are published separately.[1] For the Advanced Examination the total gross working time is 170 minutes. In their directions for administering, the authors suggest that the test be administered in four separate sittings.

Most of the separate tests have been described briefly in this chapter and in Chapter VII. However, certain important features common to all the tests or characteristic of the battery as a whole may be mentioned here. (1) The scores obtained for the different tests are equated to each other, making any given score, such as 50, signify equal attainment in each test. This of course facilitates the comparison of scores in one test with those in another. It also makes possible a valid composite score in which no one test is given undue weight. (2) The administration of the test is simple and convenient. (3) There are liberal time limits for the various tests, which provide for an accurate measurement of power in the subjects concerned. (4) The tests are easy to score and to transcribe. (5) A profile chart is provided that makes possible a graphical representation of the scores of each pupil in each test. On this chart the probable error of a score is indicated at various levels. As explained in Chapter IV, the probable error of a score is important in determining the veracity of a pupil's score in a given test. The chart

---

[1] The New Stanford Reading Test, New Stanford Arithmetic Test, the New Stanford Geography Test, and the New Stanford Language Usage Test are published separately.

permits the recording of subsequent test scores, thus making it easy to compare a pupil's performance at different intervals of time. Finally, the combined profile for the entire class may be used to determine whether a given subject has had too much or too little emphasis. (6) The reliabilities for the various tests are high. For the test as a whole they are as follows: [1]

Grade	II	III	IV	V	VI	VII	VIII	IX
Reliability coefficient	.95	.95	.89	.95	.95	.95	.96	.94

From the formula given in Chapter IV (page 83), it will be seen that the reliability coefficient is used to compute the probable error of a score. When the standard deviation has been obtained, the probable error of any given score is easily determined. (7) The test provides for the conversion of scores into three kinds of norms — age, grade, and educational age. These norms are easily understood and greatly facilitate the interpretation of the test. They also make it possible to interpret a pupil's performance for the entire test or for each test separately in terms of each one of these well-known concepts.

It should perhaps be explained at this time that "educational age" is analogous to "mental age," but the latter is obtained through the application of an intelligence test, while the former is obtained through the use of an educational or achievement test. Thus if a pupil's educational age is 12 years and 9 months, his achievement is that of the average pupil of the chronological age, 12 years and 9 months. In the same manner each subject test yields a subject age. Educational quotients and subject quotients may be obtained by dividing the educational and subject ages by the chron-

---

[1] T. L. Kelley, Giles M. Ruch, and Lewis M. Terman. New Stanford Achievement Test, Guide for Interpreting, page 9. World Book Company, Yonkers-on-Hudson, New York; 1929.

ological age. These quotients are interpreted in much the same manner as intelligence quotients; that is, an educational quotient or subject quotient of 100 signifies that the pupil's performance is in accordance with that of the average pupil for his age.

The ease with which the various norms may be read from the profile chart is best illustrated by quoting an illustration from the Guide for Interpreting.[1] According to this example, a pupil 13 years and 5 months old in the low section of the eighth grade made the following record in May:

Paragraph Meaning . .	98	History and Civics . . . 90
Word Meaning	100	Geography 95
Dictation	76	Physiology and Hygiene . 86
Language Usage	85	Arithmetic Reasoning . . 80
Literature	81	Arithmetic Computation . 93

The profile of this pupil's record appears on the chart reproduced in Figure 13.

The scores made by this pupil are recorded for each of the ten tests in the second column at the left. The number and name of each test is given at the top of the chart from left to right. Thus the pupil's score in Test 1 is located directly below it and to the right of the number in the third column, in light-faced type, which indicates his score, 98. In the same manner his scores in the other tests are located. The last four columns indicate — in the order named — total score, educational age, chronological age, and school grade. We see at once that this pupil has a total score of 88, an educational age of 13 years and 7 months, and a grade achievement of about 8.4.[2] When the scores earned by each pupil

[1] *Op. cit.*, pages 11 and 12.
[2] School grades are expressed in numerical values. September 15 is considered as the beginning of the school year. A pupil in Grade VII tested on this date would be recorded as being in Grade 7.0. One tenth of a grade is added for each month beyond this date.
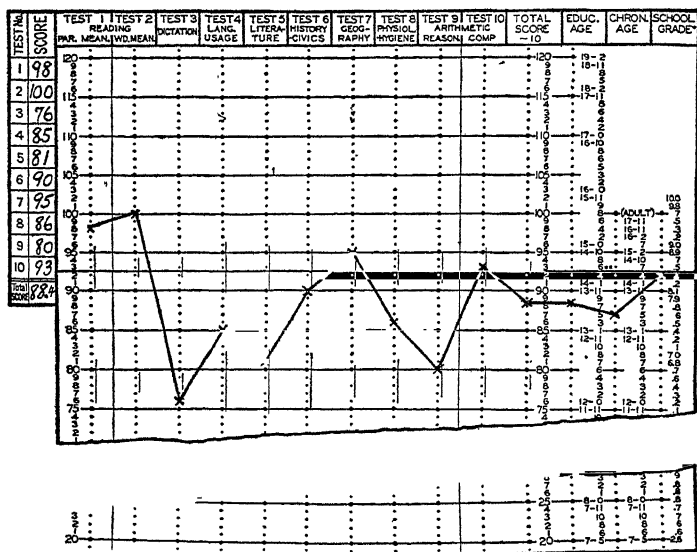
FIG. 13. Showing graphically the scores earned by a pupil in the New Stanford Achievement Test.

in a class have been recorded in this manner, the profile chart appearing on page 2 of the New Stanford Achievement Test may be detached and filed as a permanent record. In this way the score or achievement of any pupil in any subject can be found readily by the teacher or school official desiring such information.

In their directions for interpreting, the authors give many valuable suggestions for the uses and treatment of test results. Because of the ease of interpretation and the relatively high reliability of the scores in the seven important school subjects included, the New Stanford Achievement Test may well be used to classify and section pupils according to their achievement. For the same reasons it is useful in the treatment of individual cases. Owing to the high reliability of the tests,

there is relatively small danger of greatly misplacing a pupil. In this way the tests are especially valuable in properly placing pupils who come from other schools. Especially when given before entrance to the junior or senior high school, the tests may be used for the guidance and classification of pupils. A number of other uses are suggested and discussed in Chapter X.

The Stenquist Mechanical Aptitude Test. This test may be used in Grades V to VIII and in high school. The test consists of two parts or sub-tests. Test I is the simpler and includes ninety-five pictures of common mechanical objects. The pupil's score is determined by his ability to show the interrelations of the different items or to indicate how they are used. Test II is more difficult. It consists of seventy-eight pictures of mechanical objects. The pupil must show his ability in the test by classifying these objects, putting them together, or applying reasoning in some other manner. Specific directions for giving and scoring the tests are given in a manual. The scoring is made entirely objective by the use of stencils. The time allowed is forty-five and fifty minutes, respectively, for Tests I and II.

The test is intended to measure mechanical aptitude, and it has been found that it accomplishes this purpose with a fair degree of reliability. It has value as a means of prognosis for vocational training since it measures aptitude more or less irrespective of training. The test is obviously better adapted to boys than to girls. The author of the test finds that there is a low correlation between this test and general intelligence tests.[1] In other words, the scores made by a group of boys in the Stenquist Mechanical Aptitude Test are not closely parallel to those made by the same boys in a general intelligence test. Thus a pupil who scores low in

[1] J. L. Stenquist, "The Case for the Low IQ," in *Journal of Educational Research*, Vol. IV, pages 241–254; November, 1921.

general intelligence may do considerably better in the mechanical aptitude test. This fact, of course, enhances the value of the test for vocational guidance, and makes it especially useful for those high schools that provide courses along mechanical lines.

The Seashore Measures of Musical Talent. Musical talent is often considered a single trait, thus requiring but a single measure to determine to what extent it is possessed by a given individual. Dr. Seashore has shown, however, that it is a very complex capacity that may be analyzed into many somewhat independent traits. Six of these he has reduced to measurement. The test is based on six Columbia phonograph records of the standard twelve-inch type.[1] The records are played in the usual manner, both sides being used for a given test. A manual of directions provides instructions for giving the tests and for scoring and interpreting the results. For each test the pupil should have a test blank conveniently ruled for his responses. The six traits that are measured in this way are the fundamental ones underlying general musical talent: (1) sense of pitch, (2) sense of intensity, (3) sense of time, (4) sense of consonance, (5) tonal memory, (6) sense of rhythm. A pupil may obtain a very different score in each test. Since they are measures of different things, the scores may not be added in order to determine a pupil's talent. They must be interpreted as they relate to each other and to the other traits that the pupil possesses. Norms are provided to aid in this interpretation. The test has been used for surveys in elementary grades and in high school in order to locate pupils with native musical talent. Seashore gives illustrations of the results that may be attained by the discovery of gifted children who have had little or no training.

[1] Seashore's *Measures of Musical Talent*. Columbia Graphophone Company, New York; 1919.

The Kwalwasser-Ruch Test of Musical Accomplishment. This test is designed to measure all phases of a pupil's musical knowledge. It is intended for use in Grades IV to XII, and consists of ten different parts included in a printed booklet of eight pages. These parts are listed:

1. Knowledge of musical symbols and terms
2. Recognition of syllable names
3. Detection of pitch errors in a familiar melody
4. Detection of time errors in a familiar melody
5. Recognition of pitch names
6. Knowledge of time signatures
7. Knowledge of key signatures
8. Knowledge of note values
9. Knowledge of rest values
10. Recognition of familiar melodies from notation

The time limits range from three to eight minutes for each different part. All of the tests may be scored on an objective basis by means of a scoring key. The test is intended to measure achievement in public school music. It has a high reliability and so can be used to determine the proficiency of each pupil as well as the proficiency of the class as a whole. Norms for each grade are provided. The test also has some diagnostic value and may be used to locate the strong and weak points in a pupil's training.

The validity of achievement tests. The meaning and importance of the validity of a test has been discussed in Chapter IV. In selecting an achievement test for purposes of measurement, the question of validity (whether the test really measures what it purports to measure) should be given first consideration. Evidence concerning the validity of a test should be given in the manual of directions or in the other descriptive matter that accompanies the test. In other words, this information should be conveniently avail-

able to the prospective test user.   The general methods used in validating a test have been outlined in Chapter IV, but we may restate those applicable to achievement tests.

First of all the test must be representative of the subject matter on which the pupil is to be examined.   This means that the test items must constitute an adequate sampling of the various parts and topics of the subject.   Let us suppose, for example, that our pupils have studied a spelling list of a thousand words and that we desire to test their proficiency in spelling this list.   A perfectly valid measure of each pupil's achievement, as applied to this list in spelling, would be the entire list of words, given as a test.   Such a procedure would obviously consume a great deal of time. We should therefore make up a much shorter list of words for our test, and from the results obtained by its application we should infer the proficiency in spelling the thousand words. It is obvious that the care exercised in selecting the words for our test will determine its validity.   If words not in the study list are included, or if a very short list of only the easiest or the hardest words is used, it is clear that a pupil's achievement may be entirely misrepresented by the test.

There are available a number of methods of validating tests that are applicable to achievement tests.   One of these methods makes use of the combined judgments of competent persons and is illustrated in the construction of the Thorndike Handwriting Scale.   A second method requires an analysis of textbooks and courses of study in order to determine the relative importance of the various topics.   Sometimes this method is extended to include an analysis of the social utility of prospective test items.   Such an analysis may include, for example, the examination of newspapers and magazines to discover the history or geography knowledge needed by a pupil in order to read the papers intelligently.   If such socially useful data are not included in the courses of study

or in the textbooks that comprise the assigned material, they should not be included in the test. Otherwise a pupil's test score cannot be used to indicate his mastery of the subject matter that he has studied. In a third method of test validation we compute the coefficient of correlation between the test scores and school marks. The usefulness of this method is limited because of the unreliability of school marks. A fourth method requires that we compute the coefficient of correlation between the test scores and other tests already in use. The usefulness of this method will depend, of course, on the validity of the tests used as criteria.

Because each of the four methods outlined for validating tests rests fundamentally on subjective judgment, no one method may be taken as purely objective. The subjective element may be eliminated or greatly reduced, however, by careful experimental work in connection with the construction of the test. This procedure is well illustrated in the construction of the Stanford-Binet Test discussed in Chapter V. The major steps in experimental work on a test are:

1. Selection of test items for trial
2. Actual trial of the test items on the pupils of the age or grade for whom the test is intended
3. Determination of the relative difficulty of the items and their consequent weighting in the test
4. Construction of equivalent forms
5. Actual trials of each form
6. Determination of the best time limits
7. Determination of reliability by computing reliability coefficients and errors of measurement of individual scores
8. Experimental determination of directions to pupils, construction of scoring keys, etc.
9. Writing the manual of directions

The reliability of achievement tests. We have already discussed the meaning of reliability in various connections in Chapters IV, V, and VI. Chapter IV gave suggestions concerning the requisite reliabilities in testing for various purposes. We shall here present evidence pertaining to the reliability of the tests discussed in Chapters VII and VIII. Unfortunately such evidence is not always easy to find. Recent achievement tests usually give the reliability with other useful information in the manual of directions, where it can be conveniently studied by those using the tests. The data for the tests listed below were collected, together with other information about these tests, by the authors of two recent books on educational measurement.[1] While the information is not sufficient to make complete evaluations of the reliabilities of the various tests, it may be regarded as suggestive. The information is given in the parentheses following each test. The letter K indicates that it was obtained from Kelley, and the letter R that it was obtained from Ruch and Stoddard. The word " plus " means that the author considers the test more reliable than teachers' judgments, while the word " equal " means that he considers it equally reliable. The numbers are reliability coefficients.

1. Thorndike Handwriting Scale (equal K)
2. Ayres Handwriting Scale, Gettysburg (plus K)
3. Ayres Spelling Scale (plus K)
4. Iowa Spelling Scale (plus K)
5. Monroe Timed Sentence Spelling Tests (equal K)
6. Courtis Research Test in Arithmetic (plus K)
7. Woody Arithmetic Scales (.75 K)
8. Woody-McCall Mixed Fundamentals in Arithmetic (.50 to .81 K)

[1] T. L. Kelley, *Interpretation of Educational Measurements,* Chapter X. World Book Company, Yonkers-on-Hudson; 1927.

G. M. Ruch and G. D. Stoddard, *Tests and Measurements in High School Instruction,* Part II. World Book Company, Yonkers-on-Hudson, New York; 1927.

9. New Stone Reasoning Test (Reasoning, .63 to .87 *Author*) (Accuracy, .66 to .78 *Author*)
10. Monroe Reasoning Test in Arithmetic (Correct Principle, .60 K) (Correct Answer, .56 K)
11. Stevenson Problem Analysis (plus K)
12. Monroe Silent Reading Test (Rate, .75 to .83 K) (Comprehension, .65 to .66 K)
13. Thorndike-McCall Reading Scale (.56 to .80 K)(.58 to .75 R)
14. Haggerty Achievement Examination in Reading, Sigma 3 (.83 to .88 K)
15. Haggerty Achievement Examination in Reading, Sigma 1 (.88 K)
16. Charters Diagnostic Language Tests (.73 to .78 for Grade IX R) (.90 for Grades III–VIII R)
17. Wilson Language Error Test (.65 K)
18. Willing Scale for English Composition (Story Value, .40 R) (Form Value, .92 R)
19. Courtis Supervisory Geography Tests (.92 to .95 R)
20. Buckingham-Stevenson Place Geography Tests (.86 R) (.86 K)
21. Gregory-Spencer Geography Tests (.81 to .88 K)(.81 R)
22. Gregory Tests in American History (.79 R)
23. Hahn Scale for Measuring Ability in History (equal K)
24. Stenquist Mechanical Aptitude Test (Boys, .84. Boys and Girls, .97 R)
25. Seashore Music Tests (For six tests, .35 to .70 R)
26. New Stanford Achievement Test (As stated on page 190)

A List of Selected Standardized Achievement Tests

I. ARITHMETIC

Compass Diagnostic Tests in Arithmetic. For Grades 2–8. By F. B. Knight, G. M. Ruch, J. W. Studebaker, and H. A. Greene. Scott, Foresman & Co., Chicago.

Compass Survey Tests in Arithmetic. For Grades 2–8. By F. B. Knight, G. M. Ruch, J. W. Studebaker, and H. A. Greene. Scott, Foresman & Co., Chicago.

Monroe Diagnostic Arithmetic Tests. For Grades 4–8. By W. S. Monroe. Public School Publishing Company, Bloomington, Illinois.

Monroe Standardized Reasoning Tests in Arithmetic. For Grades 4–8. By W. S. Monroe. Public School Publishing Company, Bloomington, Illinois.

New Stanford Arithmetic Test. For Grades 2–9. By T. L. Kelley, G. M. Ruch, and L. M. Terman. World Book Company, Yonkers-on-Hudson, New York.

New Stone Reasoning Tests in Arithmetic. For Grades 4–9. By Cliff W. Stone. Teachers College, Columbia University, New York.

Otis Arithmetic Reasoning Test. For Grades 4–12. By A. S. Otis. World Book Company, Yonkers-on-Hudson, New York.

Public School Achievement Test in Arithmetic Reasoning. For Grades 3–8. By J. S. Orleans. Public School Publishing Company, Bloomington, Illinois.

Public School Achievement Test in Arithmetic Computation. For Grades 2–8. By J. S. Orleans. Public School Publishing Company, Bloomington, Illinois.

Schorling-Clark-Potter Arithmetic Test. For Grades 5–12. By R. Schorling, J. R. Clark, and M. A. Potter. World Book Company, Yonkers-on-Hudson, New York.

Spencer Diagnostic Arithmetic Tests. For Grades 3–8. By P. L. Spencer. Bureau of Administrative Research, University of Cincinnati, Cincinnati, Ohio.

Stevenson Arithmetic Reading Test (Problem Analysis). For Grades 4–9. By P. R. Stevenson. Public School Publishing Company, Bloomington, Illinois.

Woody Arithmetic Scales, Van Wagenen Revision. For Grades 3–8. By Clifford Woody. Public School Publishing Company, Bloomington, Illinois.

Woody-McCall Mixed Fundamentals in Arithmetic. For Grades 3–8. By Clifford Woody and W. A. McCall. Teachers College, Columbia University, New York.

II. READING

Burgess Scale for Measuring Ability in Silent Reading. For Grades 3–8. By May Ayres Burgess. Russell Sage Foundation, 130 East Twenty-Second Street, New York.

Detroit Reading Tests. For Grades 2–9. By C. M. Parker and E. A. Waterbury. World Book Company, Yonkers-on-Hudson, New York.

Detroit Word Recognition Test. For primary grades. By E. M. Oglesby. World Book Company, Yonkers-on-Hudson, New York.

Gates Graded Word Pronunciation Test (Oral). For Grades 1–8. By A. I. Gates. Teachers College, Columbia University, New York.

Gates Primary Reading Tests. For Grades 1–2. By A. I. Gates. Teachers College, Columbia University, New York.

Gates Silent Reading Tests. For Grades 3–8. By A. I. Gates. Teachers College, Columbia University, New York.

Gray Standardized Oral Reading Check Test. For Grades 1–8. By W. S. Gray. Public School Publishing Company, Bloomington, Illinois.

Haggerty Reading Examination, Sigma 1. For Grades 1-3. By M. E. Haggerty and M. E. Noonan. World Book Company, Yonkers-on-Hudson, New York.

Haggerty Reading Examination, Sigma 3. For Grades 6-12. By M. E. and L. C. Haggerty. World Book Company, Yonkers-on-Hudson, New York.

Monroe Standardized Silent Reading Tests, Revised. For Grades 3-8. By W. S. Monroe. Public School Publishing Company, Bloomington, Illinois.

New Stanford Reading Test. For Grades 2-9. By T. L. Kelley, G. M. Ruch, and L. M. Terman. World Book Company, Yonkers-on-Hudson, New York.

Public School Achievement Test in Reading. For Grades 2-8. By J. S. Orleans. Public School Publishing Company, Bloomington, Illinois.

Pressey First Grade Attainment Scale in Reading. For Grade 1. (There are similar scales for second and third grades.) By L. C. Pressey and V. Grant. Public School Publishing Company, Bloomington, Illinois.

Sangren-Woody Reading Test. For Grades 4-8. By P. V. Sangren and C. Woody. World Book Company, Yonkers-on-Hudson, New York.

Thorndike-McCall Reading Scales. For Grades 2-12. By E. L. Thorndike and W. A. McCall. Teachers College, Columbia University, New York.

III. HANDWRITING

Ayres Handwriting Scale, Gettysburg Edition. For Grades 4-8. By L. P. Ayres. Russell Sage Foundation, 130 East Twenty-Second Street, New York.

Chart for Diagnosing Faults in Handwriting. For all grades. By F. N. Freeman. Houghton Mifflin Company, Boston.

Leamer's Diagnostic Practice Sentences. For Grades 2-8. By E. W. Leamer. Public School Publishing Company, Bloomington, Illinois.

Thorndike Handwriting Scale. For Grades 2-12. By E. L. Thorndike. Teachers College, Columbia University, New York.

West Chart for Diagnosing Elements of Handwriting. By Paul V. West. Public School Publishing Company, Bloomington, Illinois.

IV. SPELLING

Buckingham Extension of the Ayres Scale. For Grades 2-8. By B. R. Buckingham. Russell Sage Foundation, 130 East Twenty-Second Street, New York.

Iowa Spelling Scale. For Grades 2-8. By E. J. Ashbaugh. Public School Publishing Company, Bloomington, Illinois.

Monroe Timed Sentence Spelling Test. For Grades 3-12. By W. S. Monroe. Public School Publishing Company, Bloomington, Illinois.

Morrison-McCall Spelling Scale. For Grades 2–8. By J. C. Morrison and W. A. McCall. World Book Company, Yonkers-on-Hudson, New York.

V. ENGLISH

Charters Diagnostic Language Tests. For Grades 3–8. By W. W. Charters. Public School Publishing Company, Bloomington, Illinois.

Charters Diagnostic Language and Grammar Test. For Grades 7–8. By W. W. Charters. Public School Publishing Company, Bloomington, Illinois.

Franseen Diagnostic Tests in Language. For Grades 3–8. By C. E. Franseen. Bureau of Administrative Research, University of Cincinnati, Cincinnati, Ohio.

Hudelson English Composition Scale. For Grades 4–12. By E. Hudelson. World Book Company, Yonkers-on-Hudson, New York.

Kirby Grammar Test. For Grades 7–12. By T. J. Kirby. Bureau of Educational Research and Service, Extension Division, University of Iowa, Iowa City.

Lewis English Composition Scales. For Grades 3–12. By E. E. Lewis. World Book Company, Yonkers-on-Hudson, New York.

Nassau County Supplement to Hillegas Scale. For Grades 4–12. By M. R. Trabue. Teachers College, Columbia University, New York.

New York English Survey Tests. For Grades 4–8. By J. S. Orleans, E. L. Cornell, W. W. Coxe, and E. B. Richards. Public School Publishing Company, Bloomington, Illinois.

New Stanford Language Usage Test. For Grades 4–9. By T. L. Kelley, G. M. Ruch, and L. M. Terman. World Book Company, Yonkers-on-Hudson, New York.

Van Wagenen English Composition Scales. For Grades 3–12. By M. J. Van Wagenen. World Book Company, Yonkers-on-Hudson, New York.

Wilson Language Error Test. For Grades 3–12. By G. M. Wilson. World Book Company, Yonkers-on-Hudson, New York.

VI. GEOGRAPHY AND HISTORY

Comprehensive Seventh and Eighth Grade History Tests. By E. C. Witham. J. L. Hammett Company, Newark, New Jersey.

Courtis Supervisory Tests in Geography. By S. A. Courtis. S. A. Courtis, Detroit, Michigan.

Gregory-Spencer Geography Tests. For Grades 6–8. By P. L. Spencer and C. A. Gregory. Bureau of Administrative Research, University of Cincinnati, Cincinnati, Ohio.

Gregory Tests in American History. For Grades 6–12. By C. A. Gregory. Bureau of Administrative Research, University of Cincinnati, Cincinnati, Ohio.

Hahn History Scales. For Grades 7–8. By H. H. Hahn. Public School Publishing Company, Bloomington, Illinois.

Hahn-Lackey Geography Scales. For Grades 4–8. By H. H. Hahn and E. E. Lackey. Public School Publishing Company, Bloomington, Illinois.

Information-Problem Test in Geography. For Grades 6–9. By B. R. Buckingham, P. R. Stevenson, D. C. Ridgley, and J. M. Shipman. Public School Publishing Company, Bloomington, Illinois.

Junior American History Test. For Grades 7–8. By H. J. Carman, T. N. Barrows, and B. D. Wood. World Book Company, Yonkers-on-Hudson, New York.

New Stanford Geography Test. For Grades 4–8. By T. L. Kelley, G. M. Ruch, and L. M. Terman. World Book Company, Yonkers-on-Hudson, New York.

Place Geography Tests. For Grades 4–8. By B. R. Buckingham and P. R. Stevenson. Public School Publishing Company, Bloomington, Illinois.

Posey-Van Wagenen Geography Scales. For Grades 5–8. By C. J. Posey and M. J. Van Wagenen. Public School Publishing Company, Bloomington, Illinois.

Pressey-Richards American History Tests. For Grades 6–12. By L. W. Pressey and R. C. Richards. Public School Publishing Company, Bloomington, Illinois.

Public School Achievement Test in Geography. For Grades 4–8. By J. S. Orleans. Public School Publishing Company, Bloomington, Illinois.

Public School Achievement Test in History. For Grades 4–8. By J. S. Orleans. Public School Publishing Company, Bloomington, Illinois.

Van Wagenen American History Scales. For Grades 5–12. By M. J. Van Wagenen, Public School Publishing Company, Bloomington, Illinois.

VII. MISCELLANEOUS TESTS

Hillbrand Sight-Singing Test. For Grades 4–6. By E. K. Hillbrand. World Book Company, Yonkers-on-Hudson, New York.

McQuarrie Mechanical Ability Test. For Grades 6–12. By T. W. McQuarrie. Teachers College, San Jose, California.

Measures of Musical Talent. For Grades 5–12. By C. E. Seashore. Columbia Graphophone Company, New York.

New Stanford Achievement Test. For Grades 2–8. By T. L. Kelley, G. M. Ruch, and L. M. Terman. World Book Company, Yonkers-on-Hudson, New York.

Public School Achievement Tests. For Grades 2–8. By J. S. Orleans. Public School Publishing Company, Bloomington, Illinois.

Stenquist Mechanical Aptitude Tests. For Grades 6–12. By J. L. Stenquist. World Book Company, Yonkers-on-Hudson, New York.

Test of Musical Accomplishment. For Grades 4–8. By J. Kwalwasser and G. M. Ruch. Bureau of Educational Research and Service, Extension Division, University of Iowa, Iowa City.

References

I. ARITHMETIC

BALLENGER, H. L. "Overcoming Some Addition Difficulties." *Journal of Educational Research,* Vol. XIII (February, 1926), pages 111–117.

BUSWELL, GUY THOMAS. *Summary of Educational Investigation Relating to Arithmetic* (Supplementary Educational Monograph, No. 27). University of Chicago; June, 1925.

—— and JOHN, LENORE. *Diagnostic Studies in Arithmetic* (Supplementary Educational Monograph, No. 30). University of Chicago; July, 1926.

HUNKINS, R. V., and BREED, F. S. "The Validity of Arithmetical Reasoning Tests." *Elementary School Journal,* Vol. XXIII (February, 1923), pages 453–466.

KELLY, F. J. "The Results of Three Types of Drill on the Fundamentals of Arithmetic." *Journal of Educational Research,* Vol. II (November, 1920), pages 693–701.

LESSENGER, W. E. "Reading Difficulties in Arithmetical Computation." *Journal of Educational Research,* Vol. XI (April, 1925), pages 287–291.

MONROE, W. S. "The Derivation of Reasoning Tests in Arithmetic." *School and Society,* Vol. VIII (September, 1918), pages 324–329.

MORTON, R. L. "Analysis of Pupils' Errors in Fractions." *Journal of Educational Research,* Vol. IX (July, 1924), pages 117–125.

OSBURNE, W. J. "A Study of the Validity of the Courtis and Studebaker Practice Tests in the Fundamentals of Arithmetic." *Journal of Educational Research,* Vol. VIII (September, 1923), pages 93–105.

SANGREN, PAUL V. "The Woody-McCall Mixed Fundamentals Test and Arithmetic Diagnosis." *Elementary School Journal,* Vol. XXIV (November, 1923), pages 206–215.

TERRY, PAUL W. "The Reading Problem in Arithmetic." *Journal of Educational Psychology,* Vol. XII (October, 1921), pages 365–377.

THORNDIKE, E. L. *The Psychology of Arithmetic.* The Macmillan Company, New York; 1922.

—— "The Constitution of Arithmetical Abilities." *Journal of Educational Psychology,* Vol. XII (January, 1921), pages 14–24.

UHL, W. L. "The Use of Standardized Material in Arithmetic for Diagnosing Pupils' Methods of Work." *Elementary School Journal,* Vol. XVIII (November, 1917), pages 215–219.

UPTON, C. B. "Influence of Standardized Tests on the Curriculum in Arithmetic." *Teachers College Record,* Vol. XXVI (April, 1925), pages 627–641.

WILSON, G. M. *A Survey of the Social and Business Usage of Arithmetic* (Contributions to Education, No. 100). Teachers College, Columbia University, New York.

II. READING

BRINKLEY, S. G. "Relative Value of Different Types of Questions in Reading Tests." *School Science and Mathematics,* Vol. XXV (October, 1925), pages 703–709.

BROWN, H. A. "The Measurement of the Efficiency of Instruction in Reading." *Elementary School Teacher,* Vol. XIV (June, 1914), pages 477–491.

CURRENT, W. F., and RUCH, G. M. "Further Studies on the Reliability of Reading Tests." *Journal of Educational Psychology,* Vol. XVII (October, 1926), pages 476–481.

DICKINSON, C. E. "A Study of the Relation of Reading Ability to Scholastic Achievement." *School Review,* Vol. XXXIII (October, 1925), pages 616–626.

GATES, ARTHUR I. "The Gates Primary Reading Tests." *Teachers College Record,* Vol. XXVIII (October, 1926), pages 146–178.

—— "A Test of Ability in the Pronunciation of Words." *Teachers College Record,* Vol. XXVI (November, 1924), pages 205–219.

—— "Methods of Constructing and Validating the Gates Reading Tests." *Teachers College Record,* Vol. XXIX (November, 1927), pages 148–159.

—— "Series of Tests for the Measurement and Diagnosis of Reading Ability in Grades 3 to 8." *Teachers College Record,* Vol. XXVII (September, 1926), pages 1–23.

GEIGER, RUTH. "A Study in Reading Diagnosis." *Journal of Educational Research,* Vol. VIII (November, 1923), pages 283–300.

GOODENOUGH, FLORENCE L. "The Reading Tests of the Stanford Achievement Scales and Other Variables." *Journal of Educational Psychology,* Vol. XVI (November, 1925), pages 523–531.

GRAY, W. S. "Case Studies of Reading Deficiencies in Junior High School." *Journal of Educational Research,* Vol. X (September, 1924), pages 132–140.

—— "Summary of Reading Investigations (July 1, 1927, to June 30, 1928)." *Elementary School Journal,* Vol. XXIX (February and March, 1929), pages 443–457, 496–509.

McCALL, W. A., and CRABBS, L. M. "Standard Test Lessons in Reading." *Teachers College Record,* Vol. XXVII (November, 1925), pages 183–191.

OGLESBY, ELIZA F. "A First-Grade Reading Test." *Journal of Educational Research,* Vol. X (June, 1924), pages 29–41.

PRESSEY, L. C. "Specific Elements Making for Proficiency in Silent Reading when General Intelligence Is Constant." *School and Society,* Vol. XXIV (November, 1926), pages 589–592.

III. HANDWRITING

BREED, F. S. "Comparative Accuracy of the Ayres Handwriting Scale, Gettysburg Edition." *Elementary School Journal,* Vol. XVIII (February, 1918), pages 458–463.

FREEMAN, FRANK N. "An Analytical Scale for the Judging of Handwriting." *Elementary School Journal,* Vol. XV (April, 1915), page 432.

—— "Some Practical Studies of Handwriting." *Elementary School Teacher,* Vol. XIV (December, 1913), pages 167–179.

KOOS, L. V. "The Determination of Ultimate Standards of Quality in Handwriting for the Public Schools." *Elementary School Journal,* Vol. XVIII (February, 1918), pages 423–446.

MORTON, R. L. "The Value of a Handwriting Scale to an Untrained Teacher." *Journal of Educational Research,* Vol. III (February, 1921), pages 133–137.

WEST, PAUL V. "Improving Handwriting through Diagnosis and Remedial Treatment." *Journal of Educational Research,* Vol. XIV (October, 1926), pages 187–198.

IV. SPELLING

ANDERSON, WILLIAM N. *Determination of a Spelling Vocabulary Based upon Written Correspondence* (Studies in Education: First Series, No. 52: Vol. II, No. 1). University of Iowa, Iowa City; 1921.

ASHBAUGH, E. J. *The Iowa Spelling Scales: The Story of Their Derivation* (Journal of Educational Research Monograph, No. 3). Public School Publishing Company, Bloomington, Illinois; 1922.

AYRES, L. P. *A Measuring Scale for Ability in Spelling.* Division of Education, Russell Sage Foundation, New York; 1915.

—— *The Spelling Vocabularies of Personal and Business Letters.* Division of Education, Russell Sage Foundation, New York; 1913.

BREED, F. S. "What Words Should Children Be Taught to Spell?" *Elementary School Journal,* Vol. XXVI (October, November, and December, 1925), pages 118–131.

DAVIS, GEORGIA. "Remedial Work in Spelling." *Elementary School Journal,* Vol. XXVII (April, 1927), pages 615–626.

GATES, A. I., and CHASE, E. H. "The Methods and Theories of Learning to Spell, Tested by Studies of Deaf Children." *Journal of Educational Psychology,* Vol. XVII (May, 1926), pages 289–300.

KALLOM, ARTHUR W. "Some Causes of Misspellings." *Journal of Educational Psychology,* Vol. VIII (September, 1917), pages 391–406.

KINGSLEY, JOHN H. "The Test-Study Method versus the Study-Test Method in Spelling." *Elementary School Journal,* Vol. XXIV (October, 1923), pages 126–129.

LESTER, JOHN A. "Spelling Ability and Meaning of Vocabulary as Indications of Other Abilities." *Journal of Educational Psychology*, Vol. XVI (March, 1925), pages 175–181.

McKEE, GRACE M. "Children's Themes as a Source of Spelling Vocabulary." *Elementary School Journal*, Vol. XXV (November, 1924), pages 197–207.

MORTON, R. L. "The Reliability of Measurement in Spelling." *Journal of Educational Method*, Vol. III (April, 1924), pages 321–328.

THORNDIKE, E. L. "The Vocabularies of School Pupils." *Contributions to Education*, Vol. I, pages 69–76. World Book Company, Yonkers-on-Hudson, New York; 1924.

WASHBURNE, CARLETON W. "A Spelling Curriculum Based on Research." *Elementary School Journal*, Vol. XXIII (June, 1923), pages 751–762.

WITTY, PAUL A. "Diagnosis and Remedial Treatment of Poor Spellers." *Journal of Educational Research*, Vol. XIII (January, 1926), pages 39–44.

WOODY, CLIFFORD. "Application of Scientific Method in Evaluating the Subject Matter of Spellers." *Journal of Educational Research*, Vol. I (February, 1920), pages 119–128.

V. ENGLISH

ASHBAUGH, E. J. "The Measurement of Language. What Is Measured and Its Significance." *Journal of Educational Research*, Vol. IV (June, 1921), pages 32–39.

BREED, F. S., and FROSTIC, F. W. "Scale for Measuring the General Merit of English Composition in the Sixth Grade." *Elementary School Journal*, Vol. XVII (January, 1917), pages 307–325.

CHARTERS, W. W. "Constructing a Language and Grammar Scale." *Journal of Educational Research*, Vol. I (April, 1920), pages 249–257.

HUDELSON, EARL. *English Composition: Its Aims, Methods, and Measurement* (The Twenty-Second Yearbook of the National Society for the Study of Education, Part I). Public School Publishing Company, Bloomington, Illinois; 1923.

LYMAN, R. L. "A Study of Twenty-Four Recent Seventh and Eighth Grade Language Tests." *Elementary School Journal*, Vol. XXIV (February, 1924), pages 440–452.

McPHEE, CLARE. "The Teaching of Language Forms." *Elementary School Journal*, Vol. XXVI (October, 1925), pages 137–146.

WILSON, G. M. "Language Error Tests." *Journal of Educational Psychology*, Vol. XIII (October, 1922), pages 430–437.

—— "Locating the Language Errors of Children." *Elementary School Journal*, Vol. XXI (December, 1920), pages 290–296.

WILLING, MATTHEW H. "Individual Diagnosis in Written Composition." *Journal of Educational Research*, Vol. XIII (February, 1926), pages 77–89.

VI. HISTORY AND GEOGRAPHY

BARTHELMESS, HARRIET M. "Geography Testing in Boston." *Journal of Educational Research*, Vol. II (November, 1920), pages 701–712.

BRANOM, M. E. "Objective Measurement of Problem Geography." *Journal of Geography*, Vol. XXV (February, 1926), pages 52–59.

BUCKINGHAM, B. R. "Correlation between Ability to Think and Ability to Remember with Special Reference to United States History." *School and Society*, Vol. V (April, 1917), pages 443–449.

—— "A Proposed Index of Efficiency in Teaching United States History." *Journal of Educational Research*, Vol. I (March, 1920), pages 161–172.

GREGORY, C. A., and SPENCER, PETER L. "A Geography Test for the Sixth, Seventh, and Eighth Grades." *School and Society*, Vol. XV (April, 1922), pages 452–456.

GOLD, M. S. "Testing Vocabulary in History." *Historical Outlook*, Vol. XVII (October, 1926), pages 285–291.

KEPNER, P. T. "A Survey of the Test Movement in History." *Journal of Educational Research*, Vol. VII (April, 1923), pages 309–325.

STEVENSON, P. R. "A Problem Test in Geography." *Journal of Educational Research*, Vol. V (April, 1922), pages 350–353.

VAN WAGENEN, MARVIN J. "Some Implications of the Revised Van Wagenen History Scales." *Teachers College Record*, Vol. XXVII (October, 1925), pages 142–149.

WASHBURNE, CARLETON, and PENDLETON, CHARLES. "The Fact Basis of a History, Geography, and Civics Curriculum." *Journal of Educational Research*, Vol. VIII (October, 1923), pages 233–238.

WITHAM, E. C. "Standard Geography Tests." *American School Board Journal*, Vol. LXXI (November, 1925), pages 51–52.

VII. OTHER REFERENCES

BELL, J. CARLETON. "Mechanical Aptitude and Intelligence." *Contributions to Education*, Vol. I, pages 270–282. World Book Company, Yonkers-on-Hudson, New York; 1924.

GIBSON, K. "Experiment in Measuring Results of Fifth Grade Class Visits to an Art Museum." *School and Society*, Vol. XXI (May, 1925), pages 658–662.

KARWOSKI, T. F., and CHRISTENSON, E. O. "A Test for Art Appreciation." *Journal of Educational Psychology*, Vol. XVII (March, 1926), pages 187–194.

SCHOEN, MAX. "Recent Literature on the Psychology of the Musician." *Psychological Bulletin*, Vol. XVIII (September, 1921), pages 483–489.

SEASHORE, C. E. "Avocational Guidance in Music." *Journal of Applied Psychology*, Vol. I (March, 1917), pages 342–348.

SEASHORE, C. E.   *The Psychology of Musical Talent.*   Silver, Burdett & Co., New York; 1919.

—— *A Survey of Musical Talent in the Public Schools* (Studies in Child Welfare, Vol. I, No. 2).   University of Iowa, Iowa City; 1920.

TOOPS, H. A.   *Tests for Vocational Guidance of Children Thirteen to Sixteen* (Contributions to Education, No. 136).   Teachers College, Columbia University, New York; 1924.

TRABUE, M. R.   "Scales for Measuring Judgment of Orchestral Music."   *Journal of Educational Psychology,* Vol. XIV (December, 1923), pages 545–561.

CHAPTER NINE

THE MEANING OF SCORES

Point scores. A point score may be defined as a score that is derived directly from the test used. It may be derived in several different ways, such as (1) counting the number of exercises correctly completed, (2) multiplying or dividing the number done correctly by a predetermined number, (3) correcting for chance where there are given to each question only two or three answers, of which the pupil is to choose the correct one. If there are only two choices, the method usually followed is to subtract the number wrong from the number right $(R - W)$. If there are three choices, the method is to subtract one half the number wrong from the number right $\left(R - \dfrac{W}{2}\right)$. If there are more than three choices, the element of chance is usually ignored. It is clear that such point scores taken by themselves do not have any meaning, but only in relation to point scores earned by a group of pupils. To illustrate, let us suppose that near the beginning of the school year a sixth-grade pupil has taken a standardized arithmetic test, such as the Woody-McCall Mixed Fundamentals, Form 1, and has made a point score of 22; that is, he has correctly computed twenty-two of the thirty-five exercises of which the test is composed. Let us further suppose that there are available no other scores with which to compare his score. The question then arises: How shall we interpret his score? That is: Is it about average for his age or grade? Is it below or above average and, if so, how much? It is evident that we cannot determine the answers to these questions in a satisfactory way without more information.

Let us next suppose that twenty-two other pupils in this grade have taken the same test at the same time and that

they have obtained the following point scores: 27, 26, 25, 25, 24, 24, 22, 22, 20, 20, 20, 19, 18, 18, 16, 16, 16, 14, 14, 13, 12, 11, 10. The median score of this distribution is 19 (located by counting in to the middlemost score). By comparing the first pupil's score in arithmetic with the median for his grade we are now able to say that this pupil is above the average for his class. However, we still do not know whether this pupil and the class of which he is a member are doing as well as might be expected of a sixth-grade class. We need norms or standards based upon the performance of other sixth-grade pupils in many schools, selected to give a typical representation of the pupils of this grade.

Grade norms. Such norms are called grade norms. Table 40 gives grade norms for each grade for the Woody-McCall Mixed Fundamentals Test. By referring to this table we find that this pupil is about average in arithmetic for a sixth-grade pupil, but that the class in which he is enrolled is, on the average, about a whole grade below the norm.

TABLE 40

WOODY-McCALL STANDARDS FOR THE BEGINNING OF THE SCHOOL YEAR
FOR FORMS I AND II

FOR THE GRADE AS A WHOLE		FOR HIGH AND LOW SECTIONS SEPARATELY			
Grade	Standard	Grade	Standard	Grade	Standard
III	6.8	III low	6.1	VI low	22.0
IV	13.1	III high	10.6	VI high	24.3
V	17.8	IV low	12.5	VII low	25.4
VI	22.5	IV high	16.4	VII high	27.4
VII	25.9	V low	17.2	VIII low	27.9
VIII	27.8	V high	19.9	VIII high	28.5

In order to facilitate the interpretation of scores in terms of grade norms, some standardized tests provide corrections that may be added to or subtracted from a given norm if only a single norm is indicated for the entire year. Thus the Woody-McCall norms listed in Table 40 require the addition of the following corrections to the norms for each month after October:

Grade	III	IV	V	VI	VII	VIII
Correction	.54	.43	.42	.24	.25	.20

A somewhat similar procedure divides each grade into a convenient number of equal parts, such as ten, and then establishes norms for each part. This, of course, facilitates the interpretation of tests by providing norms with which the scores may be compared at frequent intervals in order to determine a pupil's progress through a grade. This procedure is well illustrated in the Manual for the New Stanford Achievement Test, discussed in Chapter VIII (page 188).

A difficulty arises, however, when we consider such grade norms in connection with the elementary grades. The elementary grade subjects are not usually taught beyond the eighth grade. Consequently we are unable to determine grade norms for superior eighth-grade pupils in terms of the higher grades, because these subjects are not given in the high school grades and there are no pupils who can be tested for the purpose of establishing norms. Therefore the highest grade norm that we can establish is one based upon the average performance of pupils in the highest elementary grade, usually the eighth. The same limitations influence the determination of norms below the lowest grade in which a subject is taught. However, in some standardized tests grade norms are artificially extended. Thus the New Stanford Achievement Test provides such norms beyond the eighth grade for the elementary grade tests that it includes.

As with the lower grades, the norms are in terms of tenths; as, Grade 9.0, 9.1, 9.2, 9.3, etc. There is no serious objection to this arrangement provided we remember that it is an artificial device to assist in the interpretation of the scores made by pupils above the eighth-grade norm.

Percentiles. The use of percentiles is merely an extension of the principle explained in Chapter IV; that is, the computation of medians and quartiles. Table 19 (page 64) illustrates how percentiles may be used in the interpretation of scores. Thus if a pupil's score falls at or above the 90-percentile, we may say that he is in the best, or highest, 10 per cent of his group. Similarly, if his score falls between the 80- and 90-percentile, we may say that he is in the second highest 10 per cent of his group. Similar statements may be made about other percentiles. Ordinarily it is not worth while to compute more than the percentile for every tenth, such as the 90-percentile, the 80-percentile, the 70-percentile, etc., also termed " deciles." We may note that the median is the 50-percentile, the upper quartile is the 75-percentile, and the lower quartile is the 25-percentile.

When percentile tables have been computed on the basis of results obtained from a fair random sampling of large numbers of pupils in the various grades for which a test is to be used, they greatly facilitate the interpretation of point scores. When such tables are not available, they may be constructed from results obtained locally and may be used for the interpretation of the point scores. This procedure is of course limited, because it requires considerable work in connection with the necessary calculations. In actual practice, percentiles have not been found as useful as grade norms for interpreting the scores received in standardized tests by pupils in the elementary grades.

T-scores. Certain standardized tests provide for the conversion of point scores into " T-scores." The formula for

computing the T-score involves the use of the standard devia-
tion of the group for which the scores are derived. McCall,
the first to use this score, derived it in connection with the
Thorndike-McCall Reading Scale. In deriving it, he used
twelve-year-old pupils, regardless of the grades in which they
were located. This age was selected because pupils of twelve
years form the largest age group in school and so provide a
practically unselected group. An older group, for example
the fourteen-year-old pupils, would be affected by the elimi-
nation of some of their number from school and would there-
fore be a selected group. McCall assumed in his work that
the distribution of reading ability of this twelve-year-old
group would conform to the normal curve.

We should now recall that the standard deviation (σ)
always bears certain relations to the normal curve. For
example, if 1σ is laid off to the right of the mean and a per-
pendicular is erected at the mean and at $+ 1 \sigma$, approxi-
mately 34 per cent of the distribution will be included between
the two perpendiculars. If 1σ is laid off to the left of the
mean in the same manner, another 34 per cent approximately
will be included between the mean and $- 1 \sigma$. Thus approxi-
mately 68 per cent of all measures will lie between $- 1 \sigma$
and $+ 1 \sigma$ (sigmas laid off to the left are designated as minus
and those laid off to the right as plus). In the same manner
approximately 14 per cent of all measures lie between $- 1 \sigma$
and $- 2 \sigma$ and another 14 per cent lie between $+ 1 \sigma$ and
$+ 2 \sigma$. Thus approximately 96 per cent of all measures lie
between $- 2 \sigma$ and $+ 2 \sigma$. Consequently only 2 per cent of
measures fall below $- 2 \sigma$ and another 2 per cent above $+ 2 \sigma$.
We should now recall further that in theory the normal curve
never quite reaches the base line at either end of the dis-
tribution but approaches it more and more closely. Thus to
the right of $+ 3 \sigma$ lie only 0.13 per cent of all measures, while
similarly to the left of $- 3 \sigma$ lie another 0.13 per cent. If we

extend the curve to $-5\,\sigma$ and $+5\,\sigma$, it is clear that only an infinitesimal fraction of 1 per cent will lie beyond either of these limits. This means that in a large group of twelve-year-olds there would be few with a reading ability low enough to place them below $-5\,\sigma$. This fact is of importance in establishing the " zero point," which is defined as indicating " not any of the ability " in question — in this case, reading.

Having established the upper and lower limits of ability for his group as lying between $-5\,\sigma$ and $+5\,\sigma$, McCall proceeded, as explained on page 156, to obtain a scale with 100 units. The fiftieth unit would, of course, lie halfway between $-5\,\sigma$ and $+5\,\sigma$ or would be located at the mean for the distribution. Similarly the eighty-fifth unit would lie $3.5\,\sigma$ to the right of the mean. Instead of referring to these units in terms of fractional sigmas, which would obviously be awkward, McCall designated them as T-scores. Thus a point score falling exactly at the mean would be equivalent to a T-score of 50, on the scale of 100 units, while a point score falling $3.5\,\sigma$ to the right of the mean would be equivalent to a T-score of 85. Similarly a point score falling $3.5\,\sigma$ to the left of the mean would be equivalent to a T-score of 15, since it would be that many units to the right of $-5\,\sigma$.

In practice, the conversion of point scores into T-scores is facilitated by tables that give the equivalent T-score for any given point score. While it is rather difficult for beginners to understand the derivation, T-scores have distinct values. In the first place, they make it easy to interpret a pupil's performance. A T-score of 50 is always the norm or standard for the group concerned. In the second place, these norms facilitate comparisons between scores obtained on different tests that may be stated in terms of T-scores. That is, the T-scores on one test may be directly compared with the T-scores on another, because they are theoretically equal.

One limitation of T-scores, however, is that those derived from the twelve-year-old group are not strictly comparable with those derived from other age groups. However, they may be compared within certain limits that are determined in part by the difference in the variability of the different age groups.

Ruch and Stoddard [1] have defined the T-score by the following formula:

$$T = 50 + \frac{10(X - M)}{\sigma},$$

which may be used to compute the T-scores for any given distribution where:

T is the T-score;

X is any raw point score on a test;

M is the mean (average) of the distribution of scores for the group on which the T-scores are computed;

σ is the standard deviation (sigma) of the distribution of the scores for the group on which the T-scores are computed;

10   is introduced as an arbitrary constant to eliminate decimals;

50   is introduced so that the average T-score will be 50 and the others will be above or below this point.

Age norms. Age norms have already been explained in Chapters VI and VII as they apply to intelligence tests. When such norms are determined for achievement tests, the procedure requires that the mean or median score for each age group be found. It is, of course, necessary to utilize unselected age groups as nearly as possible. Thus if the mean point score in an arithmetic test is 48 for pupils who are 10 years and 8 months old, we may say that any pupil who makes this score, regardless of his chronological age,

[1] G. M. Ruch and G. D. Stoddard, *Tests and Measurements in High School Instruction*, page 351. World Book Company, Yonkers-on-Hudson, New York; 1927.

has an " arithmetic age " of 10 years and 8 months. Similarly, if the mean point score in a reading test is 72 for pupils who are 11 years and 4 months old, we may say that any pupil who makes this score, regardless of his actual age, has a " reading age " of 11 years and 4 months. Such age norms when they are derived for a single subject are usually called " subject ages "; but when they are derived from a composite of several subjects, they are usually called " educational ages." It is obvious that subject and educational ages lend themselves to the derivation of quotients as do the mental ages derived from intelligence tests. In other words, subject age when divided by chronological age yields a " subject quotient "; while educational age divided by chronological age yields an " educational quotient." The transmutation of subject scores into age equivalents is well illustrated by the New Stanford Achievement Test, which provides on page 2 a method for summarizing and recording the various scores obtained. The point scores, as explained in Chapter VIII, are converted into equivalent ages in plotting them on the profile chart of the test booklet. When the various scores have been recorded, the cover page may be filed as a record.

We may list the following advantages of age norms in the elementary grades:

1. Age norms make it possible to interpret a pupil's performance in terms of his chronological age. When grade norms are used, this comparison is often overlooked. In other words, if a teacher finds a fifteen-year-old pupil in Grade VI who is doing average work for the grade, she is prone to consider him equal to the average pupil and to overlook the fact that he is old for his grade.
2. Age norms make it possible to discover defects in grade classification. Thus a pupil in Grade V with an educational age of twelve is obviously not properly classified.

3. Age norms make it possible to interpret achievement in terms of intelligence in a way easily comprehended by teachers, as comparisons may be made readily between mental age and attainment.

4. Age, rather than grade, usually has the same meaning in different parts of the country. The meaning of grade depends largely on the classification of pupils and on promotional schemes. For example, in a system with seven elementary grades, Grade V would not have the same meaning that it would have in a system with eight elementary grades.

Which score or norm is best? This question is partly decided when the choice of a test is made. That is, a test is not usually standardized on the basis of all the norms or scores we have mentioned. It is well to keep in mind that all scores but the original point scores are " derived " scores. As has been pointed out in an earlier chapter (page 28), errors may occur at any step in the giving and scoring of a standardized test. It follows that there is more opportunity for error when a more complicated process for obtaining derived scores is used. Regardless of this fact, the advantage of employing various derived scores in the interpretation and use of the results is so great that it often more than offsets the increased possibility of error. We may repeat here the injunction that care should be exercised in connection with scoring a test and in following the provisions for rechecking the scores to insure greater accuracy.

<div align="center">EXERCISES</div>

1. Using the corrections for the Woody-McCall Mixed Fundamentals given on page 212, what would be the grade norm for a sixth-grade pupil tested in December? February? May?

2. Using Ruch and Stoddard's formula, convert the scores in the Woody-McCall Mixed Fundamentals, given on page 216, into

T-scores. To do this it will first be necessary to find the mean and the standard deviation. (Refer to the methods illustrated in Chapter IV.)

3. Obtain copies of the Stanford Achievement Test that have been taken by a group of pupils. Proceed to score these tests according to the directions in the manual.

References

Interpretation of Test Scores (Test Service Bulletin No. 21). World Book Company, Yonkers-on-Hudson, New York.

McCall, William A. *How to Measure in Education,* Chapters VIII, IX, and X. The Macmillan Company, New York; 1922.

Monroe, W. S. *The Theory of Educational Measurement,* Chapter VII. Houghton Mifflin Company, Boston; 1923.

——; DeVoss, J. C.; and Kelly, F. J. *Educational Tests and Measurements* (Revised and enlarged), Chapter XII. Houghton Mifflin Company, Boston; 1924.

Ruch, G. M., and Stoddard, G. D. *Tests and Measurements in High School Instruction,* Chapter XIX. World Book Company, Yonkers-on-Hudson, New York; 1927.

Symonds, Percival M. *Measurement in Secondary Education,* Chapters XIII and XV. The Macmillan Company, New York; 1927.

Worlton, J. T. "The Sigma Index Score as a Measuring Unit." *Elementary School Journal,* Vol. XXX (January, 1930), pages 354-362.

CHAPTER TEN

EDUCATIONAL USES OF STANDARDIZED TESTS

I. USES OF INTELLIGENCE TESTS

Knowledge of the child. It has long been an accepted fact that good teaching requires, in addition to other things, an intimate knowledge of each pupil's potentialities and limitations. This has compelled teachers to take cognizance of pupils' physical conditions; for example, visual and auditory defects, adenoids and diseased tonsils, and general health. It has required delving into such matters as habits of cleanliness and sanitation, diet and nutrition, amusement and recreation. These matters have even been followed into the child's home by teachers, school nurses, and others. Thus it appears that obtaining a knowledge of the child's intelligence is merely an extension of the injunction to the teacher: " Know your pupils."

Opposition to the use of intelligence tests. More than the usual opposition to the use of intelligence tests for educational purposes has developed among both laymen and educators. It has been a common objection that such tests should be used only by experts and that very few teachers are qualified to administer them. No doubt it would be very desirable for all teachers to be experts in the use of intelligence tests. However, it is unreasonable to require that teachers be expert in all the functions that they perform. Indeed in such functions as those listed in the first paragraph it has been found that teachers can accomplish results that would not otherwise have been accomplished had these matters been left to school physicians, nurses, and other experts. As we shall see, there are many ways in which a teacher may use intelligence tests to great advantage without being equipped with the knowledge and skill of psychologists who are specialists in this field.

Another objection has come from those who deny that intelligence tests really measure an inborn capacity to learn, and who assume that attempts to measure intelligence lead to fatalistic and skeptical conclusions concerning the possibilities of development through education. The first part of this objection we have already discussed elsewhere (page 87). The second part merits discussion here because of evidence showing that some teachers have taken the notion that it is a waste of time to labor with mediocre or dull pupils and that the bright will not need any attention. These are views that have had no support from any authority on the measurement of mental traits. They are based on a misconception of the relative potency of nature and nurture and on a lack of understanding of their respective functions, as we shall attempt to show.

Nature determines the limits to which an individual's potentialities may be developed under the most favorable conditions (environment). Since the most favorable conditions are seldom present, there is usually opportunity for development through improvement of the environment. Furthermore, the specific uses made of native capacity are largely determined by environment. It is easy to point to many illustrations to show that this is true. We may note, for example, the differences in customs and ideals of civilized nations, such as Germany, England, France, and the United States. That most of these differences are due to different environments is evidenced by the fact that they tend to disappear in the first or second generation of immigrants in America.

Again we may note the potency of environmental stimuli in the development of moral, political, and military leaders in times of need. Sometimes an individual may appear mediocre and may remain obscure until the environment is favorable for the development of latent genius; this was

the case with General Grant, who was an unsuccessful business man before the Civil War gave him an opportunity to develop his military ability. Genius, unlike murder, will not always out, and there are doubtless many instances where it has been wasted or misdirected. It is undisputed that character, temperament, and emotional stability are definitely related to individual efficiency and that they are greatly affected by environmental influence. Thorndike [1] believes that morality is more susceptible to modification than are the more intellectual traits.

More recently Terman [2] expressed a similar view, as follows:

> Possibly we have been laying too much stress upon the mastery of the school subjects. It is possible that we have overstressed both the efficacy and the importance of educational methods and devices designed to boost achievement. I shall not be surprised if we are in time compelled to modify our educational ideals in several respects. In the first place, I think we shall ultimately come to place more emphasis than we now do upon the ethical and social ends of education, and care more than we now do about making the school a wholesome place to live. In the second place, I think that we shall in the future stress more than we now do the training of the child in attitudes and interests, as contrasted with the imparting of subject mastery. In the third place, I feel reasonably sure that in the future, instead of abandoning the measurement of general intelligence and special abilities, we shall make greater and greater use of these and kindred instruments.

Finally, it is well known that happiness is not dependent upon the possession of great gifts or of talents bestowed upon the individual by original nature. This is a question of the right philosophy of life fully as much as of the capacity to

[1] E. L. Thorndike, *Educational Psychology, Briefer Course,* page 401. Teachers College, Columbia University; 1915.
[2] L. M. Terman, "Replying to Criticisms of Part I, and Introducing Part II, of the 1928 Yearbook of the National Society for the Study of Education," in *Journal of Educational Psychology,* Vol. XIX, page 373; September, 1928.

take advantage of material things in life. Many a man has won fame or fortune without achieving happiness, and the mediocre often live more happily and fully than the gifted.

It is clear that there is abundant scope for the play of environment. It is the function of the teacher and of the school to develop each individual so that he, as well as society, will realize most fully the capacities with which nature has endowed him. When we consider the question from this point of view, there is no conflict between nature and nurture, and both the school and the teacher have great opportunities to develop individuals of all levels of original capacity. It has, of course, long been recognized that the teacher can accomplish the most effective work with a pupil if she has a fairly accurate knowledge of his capacity and achievement. It is also an accepted principle that in teaching a group better results can be obtained if the group is fairly homogeneous in capacity and achievement. Indeed, this principle forms the basis for our present grading system. As was pointed out in Chapter I, however, standardized tests have revealed the fact that there are wide differences in capacity and achievement among the pupils of any typical school grade. It is natural, therefore, that educators should resort to standardized tests in order to secure a better adjustment between the school and the capacities and achievements of its pupils.

Early methods of providing for individual differences.[1] Long before standardized tests came into use, educators had recognized that pupils varied in their ability to progress through school. As long ago as 1870 H. T. Harris, superintendent of the St. Louis, Missouri, schools, provided for semiannual promotions. Thus a failing pupil would only be required to repeat half a year instead of a whole year. This

---

[1] For a thorough discussion of methods of adapting the schools to individual differences see *Adapting the Schools to Individual Differences* (The Twenty-Fourth Yearbook of the National Society for the Study of Education, Part II). Public School Publishing Company, Bloomington, Illinois; 1925.

plan has often been followed where a school system is large enough to permit its operation. Some school systems have extended the plan to permit promotions four times during the year. According to this plan a failing pupil might be required to repeat only nine weeks of work. Another early attempt in the United States to administer a school system so as to provide for individual differences among pupils was made by Preston Search, superintendent of the Pueblo, Colorado, schools from 1888 to 1894. The essential feature of the plan was that each pupil was allowed to progress at his own rate. No special technique, however, for testing pupils or for adapting subject matter to the needs of different pupils was provided.

During 1912–1913 Frederic L. Burk began to individualize school work in the training school of the San Francisco State Normal School. Burk developed a definite technique for individual instruction and promotion. An important part of his plan was the organization of the subject matter in the form of " self-instruction bulletins." Each pupil's progress was determined by the rate at which these bulletins were mastered. Although receiving much favorable attention, this plan was not widely adopted because it seemed rather difficult to administer under average public school conditions. A number of other ingenious plans for individualizing instruction might be discussed, among them nationally and internationally famous ones such as the Cambridge and Dalton plans. While these plans all recognize that differences exist among pupils in their ability to progress through school, they do not specifically require the use of standardized tests as a basis for ability grouping. It is contended by some authorities that better results can be obtained by providing for individual differences through a more accurate grouping on the basis of intelligence test results. Such grouping would make possible not only greater

flexibility in promotion but also a higher degree of differentia-
tion between the courses of study for the groups.

The multiple-track plan. L. M. Terman[1] has advo-
cated a five-track plan based in part upon intelligence test
results. According to this plan pupils would be tested and
classified on entrance to school under five types: (1) the
gifted, (2) the bright, (3) the average, (4) the slow, and
(5) the special or very slow. The pupil would then be as-
signed to the course having subject matter suited to his
group. Provision would be made for the transfer of a pupil
from one course to another whenever it was found that he
had been wrongly classified. According to this plan there
would also be different rates of promotion for the different
courses, or tracks, as well as differences in the methods of
teaching. A five-track plan such as the above could of course
be operated practically only in a fairly large school system.

The Trinidad plan. This is essentially Terman's plan,
put into operation by Superintendent Hobart M. Corning of
the Trinidad, Colorado, schools. An outline of the plan
will show concretely its essential features. Corning rejects
educational age as a basis for classification, maintaining that
even though pupils were started at the same educational age
they would not long remain at the same stage of achievement
because some would learn more rapidly than others. Thus
frequent reclassifications would be necessary. A similar
difficulty would occur if the classification were made on the
basis of mental age, because the rate of mental growth for
different pupils is not the same. He therefore decided to
classify on the basis of mental age and intelligence quotient
combined. The pupils were located in grades according to
mental age and within a given grade on the basis of the intel-
ligence quotient. This procedure Corning called "classifying

[1] L. M. Terman, et al., *Intelligence Tests and School Reorganization*, Chap-
ter I. World Book Company, Yonkers-on-Hudson, New York; 1922.

vertically by mental age and horizontally by intelligence quotient." By it he expected to obtain rather permanent grouping, although provision was made for shifting pupils from one section to another if it was found that they were improperly classified.

Owing to administrative difficulties, Corning found it necessary to use a three-track rather than a five-track plan, forming a slow, an average, and a bright group. These groups were designated A, B, and C, respectively. The order of letters was inverted to minimize the possibility of having parents and pupils make invidious comparisons. For each group a different curriculum was worked out. The differentiation in each subject was provided for in detail. In general the plan limited the work for the A division to the minimum essentials in each subject, gave the C division an enriched curriculum and the B division an intermediate curriculum. The actual grouping of children that resulted is illustrated in Table 41, which is based upon experiment with 1102 upper-grade pupils.

After four years of the multiple-track plan as used in the

TABLE 41[1]

SHOWING PER CENT OF CHILDREN IN VARIOUS DIVISIONS AS SUGGESTED BY TERMAN AND THE ACTUAL ARRANGEMENT WORKED OUT FOR 1102 UPPER-GRADE CHILDREN IN THE TRINIDAD, COLORADO, SCHOOLS

GROUP	PER CENT SUGGESTED BY TERMAN	PER CENT OF 1102 CASES IN TRINIDAD	NUMBER OF 1102 CASES IN TRINIDAD
Special	2.5	3.2	36
Slow (A)	15.0	22.9	253
Average (B) . . .	65.0	54.1	596
Bright (C) . . .	15.0	18.1	200
Gifted	2.5	1.5	17

[1] Hobart M. Corning. *After Testing — What?* page 72. Scott, Foresman & Co., Chicago; 1926. Used by permission of the publishers.

Trinidad Schools, Corning concludes that it is superior to the old mixed classification of school children. Among the administrative advantages, he lists the following: (1) More homogeneous grouping makes it possible for a teacher to handle larger classes; (2) retardation has practically been eliminated; (3) the plan is approved by teachers, pupils, and parents. Contrary to expectations, opposition from the parents of the slow or A group did not materialize. Corning explains that during the testing, which resulted in the reclassification, a publicity program was carried on through which the support of the public was enlisted. He also states that care was taken not to refer to groups as bright, average, or dull, although no attempt was made to deceive the parents.

Pros and cons of ability grouping. Considerable opposition to plans for grouping pupils on the basis of ability has developed among some educators. These educators urge that such grouping in the public schools is dangerous to democratic principles. They believe that pupils of all types should be classified together in order to develop common understanding. They fear that grouping will result in a feeling of superiority among the superior children and that this feeling will be encouraged, perhaps more or less unconsciously, by teachers and parents. They fear also that because of such grouping some teachers will have a fatalistic attitude toward the dull pupils. They believe that there is very little scientific evidence in support of ability grouping as a method of classifying pupils.

The foregoing arguments are chiefly based on theoretical considerations. They can be opposed by equally plausible theoretical considerations supported by experimental evidence. It can be argued, for example, that there is no conclusive evidence that mixed classes are good for democracy; and there is even the possibility that they are bad. Is it

not possible that the constant matching of dull pupils against bright will develop a feeling of inferiority in the former and one of superiority in the latter? Is it not also possible that the former may, because of their relatively poor showing, develop an antisocial attitude toward the school? It has frequently been pointed out that a bright pupil in a mixed group is more likely to work below his actual capacity than a dull pupil, and that this works not only injustice to the bright pupil but possible harm to society. Many educators decry the present glorification of athletics in high schools and colleges and the lack of a corresponding enthusiasm for scholarship. May we not expect that ability grouping would tend to correct this condition? Is it any more undemocratic to encourage those gifted with brain to forge as far as possible to the front than it is to encourage those gifted with brawn? Moreover, it can be argued that ability grouping makes it possible to take advantage of the long-accepted pedagogical principle that teaching should begin with what the child already knows.

It may be urged that a straightforward and honest recognition of individual differences, with the consequent superiority or inferiority of some individuals, is desirable. As a matter of fact, outside the school such recognition actually exists. People do realize distinctions in the value of occupations and vocations and in the ability of individuals to attain those more desirable. Does not the attempt to gloss over the presence of individual differences smack more of hypocrisy and demagoguery than of democracy? It is not a question whether we shall or shall not consider individual differences in our teaching, or group pupils according to ability; for we were doing both without serious opposition before standardized tests came into use. It is rather a question of how honest we dare to be in doing it scientifically and in recognizing the true basis for our classification.

The possibility of utilizing both intelligence and achievement tests for purposes of classification. In our discussion of the Trinidad plan (page 225), it was stated that Superintendent Corning rejected educational age as a basis for classification because their different rates of progress did not permit the pupils to remain long at the same stage of achievement. He overlooks the fact, however, that the pupils might be classified vertically on the basis of educational age and horizontally on the basis of educational quotient. This procedure would be identical with the one he proposes for intelligence tests. In the light of recent findings by T. L. Kelley [1] concerning the community of function between intelligence and achievement tests, we may question whether it would not be better, at least in certain grades, to use achievement tests rather than intelligence tests for the classification of pupils. Dr. Kelley [2] gives evidence to show that " on the average, in the neighborhood of .90 of the capacity measured by an all-round achievement battery score — reading, arithmetic, science, history, etc. — and of the capacity measured by a general intelligence test is one and the same." Since an achievement test gives weight to the actual achievement of pupils, this type appears to be preferable to the intelligence test for classification in the intermediate and upper grades. In the primary grades, where schooling has not had so much chance to function and for which there are not so many satisfactory achievement tests, classification could be made on the basis of intelligence test results.

[1] T. L. Kelley, *The Interpretation of Educational Measurements*, page 21. World Book Company, Yonkers-on-Hudson, New York; 1927.

[2] Kelley's statement holds true when the educational environment has been approximately the same for the pupils tested and when the intelligence test used is of the Army Alpha type. When intelligence is tested by such tests as the Stanford-Binet tests, which are less subject to the influence of schooling, the statement is not so true. This point is well illustrated in Dr. Fernald's work with non-readers, discussed on page 240.

The educational results as well as the possible effects on democracy would, of course, be the same as those noted in the discussion of the use of intelligence tests for the purpose of classification (page 227). However, the elimination of the term " intelligence " might alleviate the fears of individuals who dislike this term.

Intelligence tests and educational and vocational guidance. In a study of the intelligence of the high school students in a large city school system, the writer [1] found that there were certain important relations between intelligence and success in the various school subjects. While these relations were somewhat obscured by other factors, it became evident that intelligence, as measured by a typical group test, Army Alpha, is one of the important factors in determining the success

TABLE 42

COMPARISON OF INTELLIGENCE QUOTIENTS [2]

YEAR	SCHOOL	CHRONOLOGICAL AGE							
		13	14	15	16	17	18	19	20
Freshman . .	Central	152	129	120	104	102	94		
	Commerce	109	104	94	83	71			
Sophomore .	Central		154	140	127	117	114	110	
	Commerce		111	115	95	85	79		
Junior . . .	Central			142	141	130	115	112	110
	Commerce			105	113	105	86	75	
Senior . . .	Central				156	142	131	120	115
	Commerce				136	110	109	97	82

[1] I. N. Madsen, "The Contribution of Intelligence Tests to Educational Guidance in High School," in *The School Review*, Vol. XXX, pages 692–701; November, 1922.

[2] The IQ's in Tables 42, 43, and 44 are derived from point scores in the Army Alpha Test and are not to be confused with IQ's derived from the Binet Tests.

of high school students. Tables 42 to 47 present these relationships.

It is shown in Table 42 that there are important differences in the intelligence levels of the students attending the two high schools, Central and Commerce. Age for age and grade for grade, the students attending Central average higher than those attending Commerce. This relation becomes of increasing significance because Central High School offers primarily the classical and college preparatory subjects, while Commerce offers primarily subjects that prepare one directly for a chosen vocation. Before these students were tested, there was no systematic attempt to classify them on the basis of ability. The students classified themselves voluntarily. Table 43 gives a very good clue to their classification; for this table indicates a definite relation

TABLE 43

MEDIAN INTELLIGENCE QUOTIENTS OF FRESHMEN LISTED BY SUBJECTS

	BOYS		GIRLS	
SUBJECTS	Failed Students	Passed Students	Failed Students	Passed Students
Latin I and II	118	135	116	132
Ancient History	105	132	99	127
English	105	129	93	118
Algebra	111	123	99	124
Manual Training		112		
Mechanical Drawing . . .	108	110		
Typewriting	76	109	87	96
Woodworking	70	93		
Telegraphy	76	81		
Household Arts			94	107
Rapid Calculation			66	94
Cooking				83

found between intelligence and success in various subjects. Thus students in Latin I and II, or in other college preparatory subjects, who pass or fail possess higher IQ's than students who pass or fail in the vocational subjects.

TABLE 44

MEDIAN INTELLIGENCE QUOTIENTS AND SCHOOL MARKS

YEAR	SCHOOL	SCHOOL MARKS		IQ	
		Boys	Girls	Boys	Girls
Senior . . .	Central	81.6	85.6	135	125
	Commerce	83.9	85.2	96	102
Junior . . .	Central	79.7	81.6	143	128
	Commerce	79.6	81.3	105	97
Sophomore . .	Central	79.9	81.6	130	128
	Commerce	80.8	82.0	103	95
Freshman . .	Central	76.2	80.5	113	110
	Commerce	77.9	82.3	88	89

Table 44 shows the relation between the average school marks and intelligence. It will be seen that although the median IQ's for the Commerce students are materially lower than those for the Central students, the average school marks are about the same. In other words, as judged by the teachers, the students of the two schools are about equally successful in their respective courses. As was noted in connection with the Stenquist Mechanical Aptitude Tests,[1] a relatively low score in an intelligence test does not necessarily signify a low degree of ability in all capacities. On the contrary, the low correlation between intelligence scores and scores on the mechanical aptitude test indicates that there are

[1] See page 193.

many individuals who are mediocre in abstract intelligence but superior in mechanical aptitude. Doubtless the same statement can be made of other special aptitudes and talents. This helps to explain the success achieved by students with relatively low intelligence scores, in courses where other traits are perhaps fully as important as intelligence. This also helps us to understand the tremendous growth and popularity of schools of the vocational type. Students who did not fit into the curricula provided by the older type of high school that prepared for college found that the newer vocational high schools offered many valuable courses in which they could succeed.

The three preceding tables indicate plainly the need for educational and vocational guidance in high schools. Tables 45, 46, and 47, and Figure 14 (page 236) emphasize this need. Table 45 shows by five-year periods the rapid increase in the number of students attending high school in the United States since 1889. It has been roughly estimated that about one third of the young people of high school age are now enrolled. Thus we may expect the increase to continue far beyond the present enrollment. Table 46 shows the proportion of students in the Omaha high schools during 1920 who came from each of five vocational groups. From this table it appears that a smaller proportion of the young people of high school age are enrolled from the unskilled and the semiskilled groups than from the other groups. Hence these are the groups that we must expect will provide the majority of additional high school students.[1] Table 47 shows that distinct differences are represented in the intelligence levels of the five vocational groups, and that on the average, the

[1] According to the United States census for 1920 for Spokane, a typical American city, 9 and 41 per cent of gainfully employed individuals are classifiable as unskilled and semiskilled respectively. Compare this total of 50 per cent with the total of 12.6 per cent of high school students coming from this group, as shown in Table 46.

unskilled and the semiskilled groups rank lowest.   Hence we must expect to provide in the future for an even larger proportion of high school students who are not equipped to meet the requirements for college preparatory courses. This provision will require that greater care be exercised in planning the various curricula as well as in guiding pupils into the curricula that best fit their needs.

TABLE 45

INCREASE IN HIGH SCHOOL ENROLLMENT BY FIVE-YEAR PERIODS

YEAR	ENROLLMENT	PER CENT ENROLLMENT IS OF POPULATION
1889–1890	367,003	0.59
1894–1895	539,112	0.79
1899–1900	719,241	0.95
1904–1905	780,000	1.03
1909–1910	1,100,000	1.10
1914–1915	1,400,000	1.50
1919–1920	2,067,105	1.80
1924–1925	3,389,878	2.08

TABLE 46

DISTRIBUTION OF STUDENTS BY OCCUPATIONAL GROUPS

CLASS	UNSKILLED LABOR		SEMI-SKILLED LABOR		SKILLED LABOR		BUSINESS		PROFES-SIONS		TOTAL	
	Num-ber	Per-cent-age	Num-ber	Per-cent-age	Num-ber	Per-cent-age	Num-ber	Per-cent-age	Num-ber	Per-cent-age	Num-ber	Per-cent-age
Fresh-man	124	8.6	132	9.2	392	28.0	604	42.5	168	11.7	1,420	100
Soph-omore	40	4.7	45	5.2	265	31 0	376	44.3	127	14.8	853	100
Junior	23	3 5	39	6.0	152	23.0	315	47.5	130	20.0	659	100
Senior	8	1.6	21	4.4	104	20.8	232	46.2	135	27.0	500	100
Total	195	5.7	237	6.9	913	26.5	1,527	44.5	560	16 4	3,432	100

TABLE 47

MEDIAN ARMY ALPHA INTELLIGENCE SCORES BY OCCUPATIONAL GROUPS
IN OMAHA HIGH SCHOOLS

CLASS	GROUP					AGE STANDARDS	MEDIAN AGE
	I	II	III	IV	V		
Freshman .	78.8	83.5	93.0	100.6	115.7	92	15–1
Sophomore .	81.2	97.1	102.9	113.2	129.4	96	16–0
Junior . . .	98.7	94.3	112.8	119.5	138.1	101	16–10
Senior . . .	120.0	101.2	120.2	125.0	146.2	107	17–11

MEDIAN INTELLIGENCE QUOTIENTS BY OCCUPATIONAL GROUPS *

CLASS	GROUP				
	I	II	III	IV	V
Freshman . .	86	91	101	110	126
Sophomore .	84	101	108	118	135
Junior . . .	97	93	111	119	137
Senior . . .	113	95	114	118	137

* The intelligence quotients in the lower part of the table are derived from the data given in the upper part.

Figure 14 indicates the necessity for guidance from another point of view. This figure indicates the vocational choices of high school students as compared with the actual opportunities offered for entering each of these vocations. The percentages of choice are based upon statements of intended vocation obtained by Proctor [1] from 930 high school students. The percentages of opportunities are based upon the occupations of the fathers of 3432 Omaha high school students. It

[1] W. M. Proctor, *Use of Psychological Tests in the Educational and Vocational Guidance of High School Pupils* (Journal of Educational Research Monograph No. 1). Public School Publishing Company, Bloomington, Illinois; 1921.
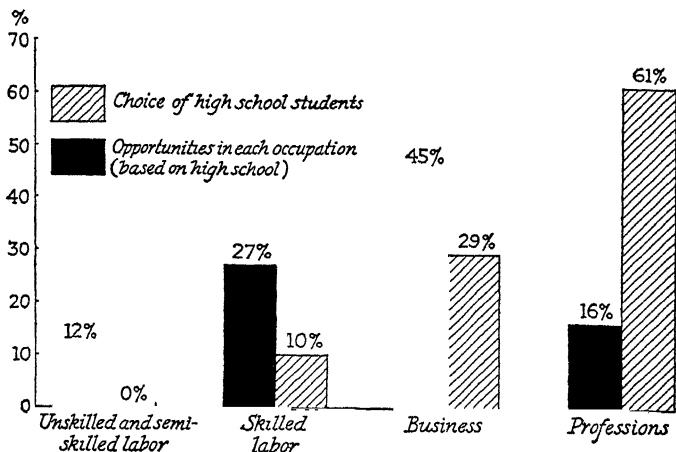
Fig. 14.   Comparison of vocational opportunities with vocational
ambitions.[1]

is clear that there is a discrepancy between vocational oppor-
tunities and vocational ambitions.   This discrepancy is so
great that many students must change their plans, perhaps
after having spent valuable time in pursuing the wrong
courses.   It should be one of the functions of educational and
vocational guidance to reduce the amount of wasted energy
indicated in Figure 14.

A comprehensive guidance program will, of course, require
much more information than is supplied by intelligence test
results.   Guidance will be made more specific and certain
with the increased accessibility of more reliable tests of special
aptitudes, emotional and temperamental traits, character
traits, etc.   With the differentiation of curricula that now
exists in the junior high school grades the guidance program
may be safely begun as early as Grades VII and VIII.   In

[1] I. N. Madsen, " The Contribution of Intelligence Tests to Educational
Guidance in High School," in *The School Review*, Vol. XXX, page 694;
November, 1922.

a study made by the writer, it is suggested that the follow-
ing data about a pupil should be made available for those
whose duty it is to give advice or guidance to pupils entering
high school :

    1. Name of pupil
    2. Chronological age
    3. Intelligence quotient
    4. Mental age
    5. Special talents
        a. Musical
        b. Mechanical
        c. Others
    6. Emotional and temperamental traits
    7. Previous school success
    8. Vocational ambitions
    9. Economic and social status [1]

II. USES OF ACHIEVEMENT TESTS

The educational problems that are solved by the aid of
achievement tests may be classified as instructional, super-
visory, and administrative. Because these groups overlap a
great deal, however, no attempt is made here to list specific
problems according to this classification. The more common
problems in the solution of which achievement tests will be
found helpful are those pertaining to the following :

    1. Diagnosis
    2. Comparison with norms
    3. Educational and vocational guidance
    4. Promotion and classification
    5. Motivation
    6. Research
    7. School reports and records

Diagnosis. The function of diagnosis, common to many
standardized tests, has already been discussed in some detail

[1] *Op. cit.,* page 701.

in connection with diagnostic tests. Diagnosis may be made either from the standpoint of the classroom teacher or from that of the supervisor. The classroom teacher is interested in discovering the specific disabilities of a pupil in each subject with a view to taking remedial steps. To be complete, therefore, such diagnosis should not only locate actual disabilities of performance but should suggest causes for the disabilities. It is also of interest to the classroom teacher to compare a pupil's standing with that of the rest of his class or with the norms for his class. Such a comparison is especially significant in discovering whether the pupil is working up to his mental capacity. To make this comparison, the achievement test, or tests, should be supplemented by a mental test. A procedure for using test results to make this comparison — which has been called the " AQ procedure " — will be discussed later in this chapter. To the supervisor diagnosis is also important in answering the problem of whether the various subjects are receiving the proper emphasis.

Comparison with norms. Standardized tests make it possible to determine the general level reached by individual pupils and the average level reached by classes or by schools. The first is of special interest to the teacher, and the last two are of interest to both the teacher and the supervisor or administrator. The teacher needs to know to what extent the results obtained in her class vary from the generally accepted norms in the different subjects. Without making such necessary comparisons, it is possible that she may overemphasize the subjects that she prefers to teach and neglect those she dislikes. The supervisor must know how effectively the work of the teacher is being done. Here again the intelligence of the pupils must be considered, for it is not reasonable to expect a teacher with a dull class to obtain results as good as those obtained by one with a superior class.

Educational and vocational guidance. Achievement tests may be used to supplement intelligence and aptitude tests in connection with educational and vocational guidance. They are especially important in the junior high school grades, where there is a wider range of choice. It has been a common belief among teachers that an attempt should be made to raise all pupils to the same level of achievement in all subjects. Kelley,[1] however, points out that there is a strong possibility that inequalities of performance in school subjects may have their roots in original nature and that these " idiosyncrasies " should be recognized in educational and vocational guidance. It will not always be easy, of course, to determine whether these inequalities in performance are caused by corresponding differences in the original nature of pupils, or whether they are caused by differences in the efficiency of teaching. If a general testing program shows that a class as a unit tends to test materially higher in some subjects than in others, we may conclude that a proportion of the inequalities that appear are due to differences in the degree of emphasis that is placed on the teaching of different subjects. Such differences can be corrected by shifting the emphasis placed on the various subjects. Kelley gives a minimal list of ten traits that should be studied for the understanding of typical school children. The following is a summary statement [2] of these traits:

1. Name, sex, and other traits, mainly those of a physical and personal nature
2. Maturity, or present chronological age
3. Verbal intelligence (as measured by achievement tests of the reading and vocabulary type as well as by general intelligence tests)

[1] T. L. Kelley, *Interpretation of Educational Measurements*, **Chapter V.** World Book Company, Yonkers-on-Hudson, New York; 1927.
[2] The student should refer to Kelley for the complete statement.

4. Social intelligence (as estimated by teachers and others who come into intimate contact with the child)
5. Activity and mechanical intelligence
6. Interests indicated by (3), (4), and (5) or otherwise specially exhibited by the child
7. Ability with reference to quantitative phenomena — computation, etc.   (Can be determined by computation and various other number tests)
8. Ability with reference to spatial relationships — geometrical forms, etc.   (Can be determined by form boards, geometrical form tests, etc.)
9. Memory with reference to verbal material
10. Special sensory or motor interests and abilities (Certain of these: readily ascertainable by existing objective tests)

Promotion and classification. The possibility of using achievement tests for the classification of school children has already been pointed out in the section on intelligence tests (page 229).   Achievement tests may also be used to place new pupils coming from other schools and to select pupils for special classes.   Thus Sutherland[1] has shown that many pupils who are normal or superior in intelligence may have in one or more subjects special disabilities that can be removed in a comparatively short time by placing the pupils temporarily in adjustment classes.   Similarly Fernald[2] discovered pupils of normal mental ability who had not learned to read or spell after several years in school.   In a group of seven of these pupils who could not read, one had an IQ of over 140.   Fernald reported that after six months of remedial treatment all but two of the seven were restored to the regular grades corresponding to their chronological ages.   A number

[1] L. M. Terman et al., *Intelligence Tests and School Reorganization*, Chapter III.   World Book Company, Yonkers-on-Hudson, New York; 1922.
[2] *Ibid.*, Chapter VI.

of other investigators have obtained results similar to those of Sutherland and Fernald. Thus Gray [1] has shown that many pupils are deficient in silent reading, not because of defective mentality, but because of disabilities in this subject that can be removed.

With the exception of the placement of pupils from other schools the uses of tests suggested in the preceding paragraph are limited to the larger school systems. The smaller schools may, however, utilize the tests for making special promotions, general reclassification, etc. To a large extent, differences among pupils in the smaller schools must of necessity be provided for by some plan of individual instruction. A noteworthy example in which achievement tests are used in connection with individual instruction is the Winnetka plan as developed by Washburne. Since this plan is not limited to large school systems but can also be operated in the smaller schools, we shall describe it in some detail.

The Winnetka plan. This plan is really a modern adaptation of Burk's plan, described earlier in this chapter (page 224). It was introduced into the schools of Winnetka, Illinois, by Washburne during 1919. An important feature of the Winnetka plan is the use of tests that cover fully the subject in which the pupils are tested. The tests are used by the teachers to diagnose the individual difficulties of the pupils. Since there were but few satisfactory tests available for this purpose, Washburne and his teachers proceeded to devise their own. In general the tests resemble the Compass Diagnostic Tests discussed in Chapter VIII in that they attempt to cover each subject completely. Practice material corresponding to each topic or process in the test

[1] W. S. Gray, "Individual Difficulties in Silent Reading in the Fourth, Fifth, and Sixth Grades," in *Report of the Society's Committee on Silent Reading* (The Twentieth Yearbook of the National Society for the Study of Education, Part II), pages 39–53. Public School Publishing Company, Bloomington, Illinois; 1921.

also was devised so that a pupil could practice on his specific deficiencies as they were revealed by the test. The attempt was to make the practice material as nearly self-instructive as possible in order to economize the teacher's time.

In addition to this systematic use of tests, the Winnetka plan provides for definite "achievement units" of work in the drill, or tool, subjects. Each work unit must be satisfactorily completed before a pupil is' allowed to proceed to another.

> Children's marks and promotions are based entirely upon individual work. There are no recitations; there is no repetition; there are no failures; there is no skipping. No child is held back to a slower rate of progress than is natural to him; none is forced too rapidly for thorough work. Each "goal" must be achieved by each pupil before he can go to the next goal.[1]

In addition to the work units described above, the Winnetka plan prescribes "social work" requiring that pupils enter into group undertakings and providing opportunities for coöperation. Activities, such as dramatizations, group games on the playground, assemblies, group singing, and the like, are included in this social work, and occupy from one third to one half of the total school time. In addition to these activities the pupils do the editing, typesetting, proofreading, and business managing of their school paper. In all this procedure there is no recitation, no marking or promotion. The general aims are to encourage self-expression and group coöperation. Washburne says:[2]

> The social part of our work is not measured. It consists of that range of subject matter to which we wish to expose our children, but which they need not master, and of opportunities to enter into undertakings which will develop children's ability to coöperate with one another.

[1] Carleton Washburne, "Educational Measurement as a Key to Individual Instruction and Promotions," in *Journal of Educational Research,* Vol. V, pages 195–206; March, 1922.

[2] *Op. cit.,* page 198.

In another place, however, he writes: [1]

> The report card used in the Winnetka schools shows on one side the pupil's progress in mastering commonly needed knowledges and skills, on the other side his progress in developing certain attitudes and habits essential for social living.

Since the latter of the two statements quoted from Washburne is the more recent, it probably indicates his present practices with regard to measuring. Certainly it is difficult to understand how any school activity can be justified or how progress in it can be determined without some method of evaluation and measurement. Incidentally we may note that the Winnetka plan, in its attempt to develop the moral and social traits of children, is in harmony with the views of both Thorndike and Terman on this matter — which views are quoted earlier in this chapter.

As the proof of the pudding is in the eating, so the test of any system of school organization is in the results it produces. Washburne has investigated such questions as the following in connection with the Winnetka plan: [2]

1. Does the individual pupil save time?
2. Does individual work increase or decrease socialized and self-expressive activities?
3. Does individual work put children through school too fast?
4. Is individual work more effective, or less, than class instruction in teaching tool subjects?
5. Does individual instruction cost more than class instruction?

---

[1] Carleton Washburne, "The Philosophy of the Winnetka Curriculum," in *Curriculum-Making: Past and Present* (The Twenty-Sixth Yearbook of the National Society for the Study of Education, Part I), page 225. Public School Publishing Company, Bloomington, Illinois; 1927.

[2] This evidence as well as evidence relative to the success of other plans of individual instruction is presented in the Twenty-Fourth Yearbook of the National Society for the Study of Education referred to earlier in this chapter.

6. Does individual instruction place too heavy a burden on the teacher?

7. How does individual work in the elementary school affect pupils' efficiency in the high school?

Space does not permit the presentation of Washburne's evidence here. On the whole, however, it seems to support the Winnetka plan.

Motivation through standard tests. Starch [1] quotes results obtained in a certain elementary school to show how a definite knowledge of success affects the progress of pupils. In this school standardized tests were given each month in reading, writing, spelling, and arithmetic in order to determine the progress that had been made. Each pupil observed his own record from month to month and made comparisons with his previous records. The results showed that the pupils made an average gain in some of these subjects twice as great as that ordinarily made in a year. Starch thinks that this gain cannot be attributed to familiarity with the material used, as new material was employed each time.

Similar results are shown by Monroe [2] in connection with the use of the Monroe Silent Reading Tests. After citing experimental results, Monroe concludes that the gain is greater than normal for the period between February and June; and in some of the grades is even four or five times as great. After allowing for the gain due to an increased acquaintance with the tests, Monroe concludes: " The adoption of the norms for these tests as educational objectives greatly motivated the school work, so that unusual progress was made by the pupils in the field of silent reading."

Diagnostic tests are particularly useful to give the pupil definite information of his success or failure in acquiring

[1] Daniel Starch, *Educational Psychology*, pages 177–178. The Macmillan Company, New York; 1927.

[2] W. S. Monroe, *Theory of Educational Measurements*, pages 252–256. Houghton Mifflin Company, Boston; 1923.

skill or information in a given subject or topic. Starch [1] says in this connection : " Much experimental work implies that the feeling of satisfaction resulting from successful trials of a task facilitates the formation of the connections concerned. It seems obvious therefore as a practical matter that precise knowledge of the success or failure on the part of the learner is exceedingly important." Gates [2] similarly states : " The abridgment of trial-and-error learning is greatly facilitated by preventing and correcting errors." From these quotations it is clear that the use of diagnostic tests in connection with teaching is considered by eminent authorities to be based upon sound psychology.

The relation between standardized tests and research and experiment. Tests may be used by teachers, supervisors, and administrators for research and experimentation. Some types of tests by their very nature imply research functions. This is true, for example, of diagnostic tests, whose outstanding function is the discovery to the pupil, as well as to the teacher, of weaknesses and strengths found in a given topic or subject. Other tests may yield valuable information as a by-product of the main purposes for using them. For example, while conducting a testing program in arithmetic in Grades IV to VII, the writer found that many teachers had the notion that it would not be fair to compare the results obtained in testing pupils during the early morning hours with those obtained in testing during the later hours of the day. The writer, therefore, had each pupil record the hour when he was tested, together with such information as name, age, grade, etc. This made possible a comparison of the results obtained from pupils tested during different hours of the day. It appeared that there was no significant difference.

[1] *Op. cit.,* page 177.
[2] Arthur I. Gates, *Psychology for Students of Education,* pages 276–277. The Macmillan Company, New York; 1923.

Frequently teachers can compare different methods of teaching a subject, particularly if there is more than one division of the same class. This not only relieves the monotony of teaching but tends to keep the teacher from falling into a rut in her teaching. The more formal and extensive experiments require the coöperation of teachers, supervisors, and administrators, and may also make it necessary to prolong the period of experimentation. For example, the problem referred to regarding sectioning classes into homogeneous groups on the basis of intelligence could be subjected to experiment that would require a year or more. Such projects should, of course, be carefully planned and executed. While teachers, supervisors, and administrators can all participate in such an undertaking, the general direction of it should be given to one individual or to a few persons who are familiar with the requirements of scientific research.

School records and reports. One valuable use of standardized tests will be neglected unless some systematic cumulative record is kept for each pupil tested. A difficulty encountered in keeping a record comes from the fact that few tests are in terms of the same units of measurement. This makes it necessary, of course, to record with each score the norm for the particular test used. If all tests could be stated in scores having the same meaning, the problem would be greatly simplified. Such scores as the T-scores and the Stanford Achievement Test scores discussed in Chapter IX come nearest to stating the different tests in terms of the same unit. Comprehensive test batteries, such as the Stanford Achievement Tests, usually offer convenient methods for recording and filing scores. For example, as has already been pointed out (page 192), the cover page of the Stanford Achievement Test may be used as a permanent record of each pupil's performance and may be detached and kept on file for this purpose.

Numerous devices have been employed for recording, graphically or otherwise, results obtained from standardized tests. It has been a rather common method to state the results in terms of grade norms because teachers are accustomed to think of pupils in relation to their grade placement. The examples shown in Figures 15 and 16 will serve to illustrate this method. The type or form of record card for different schools should be worked out exactly to fit the testing program that the particular school has outlined for itself. A record card similar to that in Figure 15 could be considerably expanded by providing for other data about the pupil or the tests given — such as names of tests used, attendance and health records of pupils, results for tests at other periods, etc.
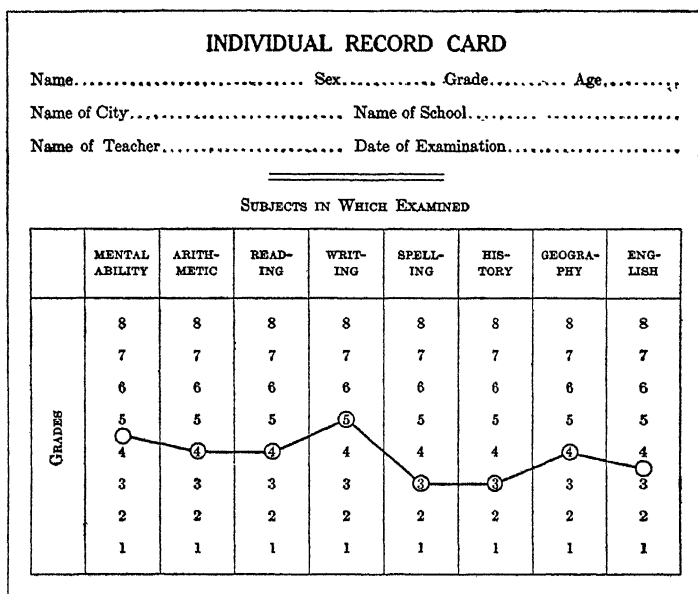
FIG. 15.   Individual record card showing the results of testing.

Name of Pupil.. Date Tested ....

Date of Birth. .... Age Last Birthday................ School Grade.....................................

School.. City................................ .... State. ....

Parent's Name .... Occupation ....

A B C D E F G H I J K L M N

HIGH SCHOOL

School Grade for which Pupil's Score is Normal

EIGHTH SEVENTH SIXTH FIFTH FOURTH THIRD

SCORE

POWER IN ARITHMETIC

EDUCATIONAL ACHIEVEMENT CHART
EIGHT-GRADE-SCHOOL EDITION

Prepared and Distributed by M. R. Trabue, Director of the Bureau of Educational Research, University of North Carolina, Chapel Hill, N. C.

DIRECTIONS: *Record in the bulb of each thermometer the score obtained on the test which it represents. Then draw a bright colored line (red or blue) upward through the stem of the thermometer to the point corresponding to the size of the score. By following the horizontal dotted lines to the left, one may readily see for what part of the elementary school course the recorded score is normal. Use one chart for the scores of each pupil and one for the average scores of each class or grade. The tests represented above are as follows*

A. T Scale values on Thorndike—McCall Reading Scales.
B. ───────────────────────────────
C. ───────────────────────────────
D. 25 word, H—AF list, Buckingham—Ayres Scale.
E. Hillegas English Composition Scale.
F-I. Woody Arithmetic Scales, Series B.

J. Buckingham Scale of Reasoning Problems.
K. Thorndike Handwriting Scale.
L. ───────────────────────────────
M. ───────────────────────────────
N. Mentimeters.

FIG. 16.   Educational achievement chart to show the results of testing.[1]

Record cards such as those described, or modifications of

[1] M. R. Trabue, " A Graphic Chart for Representing Educational Achievement Scores," in *Journal of Educational Research*, Vol. IX, pages 411–414; May, 1924.

them, might also be used to report the results obtained in standardized tests to the parents. Whether it would be desirable to do so is another matter. Parents could probably be educated to an understanding of the meaning of scores in a standardized test if a simple device for reporting were used. As we shall see, however, in Chapter XI, standardized tests cannot entirely take the place of teachers' examinations. Since this is the case, it would perhaps be better to report a pupil's standing to his parents in the terms with which they are already familiar. If teachers' examinations are improved through such methods as we suggest in the following chapter, they may be used to give accurate and frequent reports to parents.

III. THE ACCOMPLISHMENT QUOTIENT

The interpretation of achievement in terms of intelligence. It was inevitable that the use of intelligence and achievement tests for testing school children should lead to a comparison of the scores made by each pupil in the two types of tests. The idea of systematically testing pupils in such a way that achievement could be interpreted in terms of intelligence apparently occurred to several investigators at about the same time. Thus Franzen [1] during 1920 and again in 1922 published the results of a study carried on for a period of two years, in which he found that bright pupils were less likely to work to capacity mentally than dull children. He attempted to determine whether it was possible to induce pupils of different levels of intelligence to work more nearly up to their mental capacity. By a method of reclassification on the basis of intelligence, he succeeded in bringing about a closer correlation between intelligence and achievement. Pintner

[1] Raymond H. Franzen, "The Accomplishment Quotient," in *Teachers College Record,* Vol. XXI, pages 432–440; November, 1920. Also *The Accomplishment Ratio* (Contributions to Education, No. 125). Teachers College, Columbia University; 1922.

and Marshall [1] during 1921 published the results of a mental-educational test survey, in which they noted discrepancies between ability to accomplish as measured by intelligence tests, and actual accomplishment as measured by educational tests. Monroe and Buckingham [2] during 1920 published a battery of tests which included an intelligence test and also several achievement tests. Their manual of directions gave instructions for the interpretation of achievement in terms of intelligence. During 1921 Madsen [3] reported a method that he had found useful in interpreting the achievement of a class in terms of their intelligence. He pointed out that unless the intelligence of a class is taken into consideration comparisons with other classes are unfair. He also pointed out the possibility of making similar interpretations of the achievement of individual pupils.

While there were some differences in details of methods for measuring the achievement of pupils in terms of their intelligence, the fundamental concepts proposed by these investigators were the same. The essential idea is to divide educational age (EA) by mental age (MA) to obtain the accomplishment quotient (AQ).[4] Thus the formula would

[1] Rudolf Pintner and Helen Marshall, "A Combined Mental-Educational Survey," in *Journal of Educational Psychology*, Vol. XII, pages 32–43; January, 1921. Also "Results of the Combined Mental-Educational Survey Tests," in *Journal of Educational Psychology*, Vol. XII, pages 82–91; February, 1921.

[2] W. S. Monroe and B. R. Buckingham, *Teachers Handbook, Illinois Examination*. University of Illinois, Urbana; July, 1920.

[3] I. N. Madsen, "Interpretation of Achievement in Terms of Intelligence," in *Educational Administration and Supervision*, Vol. VII, pages 344–347; September, 1921.

[4] The terms "achievement quotient," "accomplishment quotient," and "accomplishment ratio" have been used synonymously. The AQ may also be computed by dividing the EQ by the IQ. The equivalence of the two formulae may be seen from the following:

$$AQ = \frac{EQ}{IQ} = \frac{\dfrac{EA}{CA}}{\dfrac{MA}{CA}} = \frac{EA}{MA}$$

be AQ $= \dfrac{\text{EA}}{\text{MA}}$. To illustrate, suppose a pupil has an EA of 9 years and an MA of 10 years. His AQ would then be 90. This AQ would indicate that this particular pupil is not working up to the level of his mental ability. Similarly a pupil with an EA of 12 years and an MA of 10 years would have an AQ of 120, which would indicate that this pupil is putting forth more than the ordinary amount of effort. Franzen attempted, in his experiment, to secure classification and motivation of the type that would result in each pupil's earning an AQ of approximately 100.

Limitations of the AQ procedure. For a time the AQ was hailed as an ideal measure to determine the efficiency with which a pupil was working, as well as to determine the efficiency of teaching. It soon appeared, however, that there were certain dangers and limitations in its use. In this connection Freeman [1] points out that its validity " depends on the clearness and distinction between measures of native capacity and of training as well as upon the accuracy of measurement of both." Kelley's statement regarding the community of function between present intelligence tests and achievement tests calls attention to the same fact in another way.[2] If this community of function is as great as Kelley states (90 per cent), then the AQ, instead of being based upon all of each of the tests used, would be based upon only a small part of each. This also would greatly reduce the accuracy of the results. Even if this limitation were not present, there would still be a limitation; for even the best available intelligence tests and achievement tests are not entirely reliable. As a result the formula for the AQ consists of a fraction in which both denominator and numerator may

[1] F. N. Freeman, *Mental Tests*, pages 26 and 286–291. Houghton Mifflin Company, Boston; 1926.
[2] *Interpretation of Educational Measurements*, page 22.

be in error, which would in turn increase the error in the resulting quotient. Another limitation, pointed out by Ruch and Stoddard,[1] is the assumption that mental age is a valid index to learning capacity, and equally valid to the same extent for all subjects. The point is that other traits besides intelligence enter into a pupil's performance and that their influence is felt more in some subjects than in others. We have already seen that the correlation between intelligence and mechanical aptitude is low. Doubtless intelligence does not play so important a part in some subjects — manual training, cooking, drawing, music, etc. — as it does in others — arithmetic, reading, history, geography, etc.

The use of age scores as bases for computing AQ's has also been criticized. Ruch and Stoddard affirm that they are not reliable or satisfactory for use in high school subjects.[2] This relates to what was said in Chapter VI (page 128) about the difficulty of getting norms on which to base mental ages for the superior students of high school age. Franzen,[3] one of the most enthusiastic promoters of the AQ procedure, has himself admitted that there are objections to the use of mental and educational ages, and consequently to the use of intelligence and educational quotients, for the purpose of comparing the performance of pupils in various tests. He states that subject ages in different tests are not necessarily equivalent. As a consequence of this the various quotients obtained would not necessarily be comparable. To remedy this difficulty Franzen proposes to use "sigma indices"[4] to compare achievement in one test with achievement in an-

[1] G. M. Ruch and G. D. Stoddard, *Tests and Measurements in High School Instruction,* page 17. World Book Company, Yonkers-on-Hudson, New York; 1927.

[2] *Ibid.*

[3] Raymond H. Franzen, "Statistical Issues," in *Journal of Educational Psychology,* Vol. XV, pages 367–382; September, 1924.

[4] Sigma indices are essentially the same as T-scores.

other. The method suggested by Franzen was used by him in his survey of Contra Costa County, California. In a report of this survey [1] he gives tables showing sigma indices for the tests that he used. One of the tests used was the Thorndike-McCall Reading Test. It will be recalled that this test gives subject age norms into which the point scores can be converted. Consequently reading quotients can also be computed. This makes it possible to compare reading quotients of pupils in this test with their sigma indices in the same test. The writer had this test given to the pupils in three grades and had both reading quotients and sigma indices computed for each pupil. In addition he had reading quotients computed by a rather roundabout method that he had used before achievement tests were standardized in terms of age norms.[2] Table 48 shows the intercorrelations among these three measures derived from the point scores of the pupils in three grades who took the tests.

From these correlations it will be seen that a pupil's standing in his class is practically the same whether expressed in terms of EQ computed by either of the two methods, or in terms of sigma indices. Thus if it is desired to find to what extent a class as a whole is working to capacity, there is no advantage in using sigma indices. If it is desired, however, to compare a pupil's achievement in one subject or test with his achievement in another, the sigma indices are useful. The argument here is the same as that used in connection with T-scores. There is one erroneous assumption in the computation of sigma indices, however, which should be pointed out. This assumption is that a homogeneous age group is used. In actual practice the age group used,

[1] Raymond H. Franzen and William H. Hanlon, *The Program of Measurement in Contra Costa County.* Standard Print, Martinez, California; 1923.

[2] I. N. Madsen, "Interpretation of Achievement in Terms of Intelligence," in *Educational Administration and Supervision,* Vol. VII, pages 346-347; September, 1921.

such as the twelve-year-old group, includes a range of twelve months. Thus, according to tables given by Franzen,[1] a pupil of 8 years and 11 months with a point score of 12 in the Woody-McCall Test would have a sigma index of 103, while a pupil of 9 years and 1 month would have a sigma index of 93. Accordingly, most of the difference between 93 and 103 is spurious.

What shall be our attitude toward the AQ procedure? In spite of the foregoing limitations and dangers in the use of the AQ, the concept is a useful one, particularly for those elementary grade subjects in which learning is most closely related to general intelligence. Indeed, as long as we use intelligence tests or have the present concept of differences in brightness among school children, teachers will judge achievement in terms of intelligence. Without the use of objective tests such judgments will be about as worthless as

TABLE 48

SHOWING INTERCORRELATIONS BETWEEN THORNDIKE-McCALL EQ's, FRANZEN'S SIGMA INDICES, AND EQ's DERIVED BY MADSEN'S METHOD

GRADE	NUMBER OF PUPILS		SIGMA INDICES	EQ's BY MADSEN'S METHOD	S. D.	MEAN
V	43	Thorndike-McCall EQ	.98	.98	13.6	113.0
		Sigma Indices		.97	11.4	109.4
		EQ's (Madsen's Method)			14.5	106.0
VI	38	Thorndike-McCall EQ	.96	.97	15.1	107.6
		Sigma Indices		.97	12.0	104.7
		EQ's (Madsen's Method)			13.6	102.6
VII	86	Thorndike-McCall EQ	.96	.96	16.5	100.8
		Sigma Indices		.97	15.1	101.0
		EQ's (Madsen's Method)			17.7	102.6

[1] *The Program of Measurement in Contra Costa County*, pages 7–14.

other subjective estimates. While the AQ should be interpreted cautiously where individual pupils are concerned, it may well be used to determine general tendencies. In other words, we should expect better performance from a group of pupils whose average IQ is high than from a group whose average IQ is low. In connection with certain subjects the interpretation of achievement in terms of intelligence is useful to determine the efficiency of teaching. Other things being equal, it is good practice to attempt to get pupils to invest their whole mental capital. It should be noted that the AQ procedure is most justified when intelligence is measured by a test like the Stanford-Binet, a test relatively free from the influence of schooling as is shown by comparison with group tests of intelligence.

IV. PLANNING TESTING PROGRAMS

The preceding pages of this chapter have discussed several administrative, supervisory, and teaching problems that give rise to the need for using standardized tests. Certain of these problems recur rather regularly. It is advisable, therefore, to plan for a definite and continuous testing program, which will make the testing an integral part of the regular school work and will not disrupt or disorganize the activities of teachers and pupils. It goes without saying that there should always be a definite and important reason for testing. It should never be done from curiosity or simply because it seems the up-to-date thing to do.

Regular and continuous testing programs. In the larger school systems testing programs should be in charge of a director who has considerable expert knowledge about all phases of testing. This director should have charge of the general survey tests that are given to all the pupils in the school system or, if the survey is limited to certain grades,

to all the pupils in those grades.   Centralized control will result in the use of better tests and more uniformity and comparability in the results obtained.   It will also eliminate the danger of overusing a given test, a procedure that may cause pupils to become " test wise " because of increased familiarity with the test and may invalidate their scores, which would be too high as compared with the general norms for the test.   Unscrupulous teachers would also be tempted to obtain the tests beforehand and coach their pupils on the content.   The writer has actually encountered this condition where a certain test was repeatedly used by a supervisor at regularly recurring intervals during the school year.   Needless to say, the results so obtained are far worse than useless.   If [such results are used at all to classify pupils or to evaluate the teacher's efficiency, they will result in error and injustice.

In carrying out the program, a director will need considerable assistance.   For example, in a large school system it will not be possible for him personally to give the tests to all the pupils.   He may therefore enlist the aid of supervisors and principals to give the tests in their respective buildings.   Another procedure, somewhat less satisfactory, is to have each teacher give the tests in her own room.   In either case, adequate preliminary training should be given to the examiners.   The importance of such training has been fully discussed in Chapter II.   Whether the aid of supervisors and principals or that of teachers should be enlisted to give tests will depend partly on the chief purposes of the survey.   A small number of well-trained examiners will give better results when the purpose of testing is to obtain data to be used to compare the teaching efficiency of different teachers, to compare one method of teaching with another, or for some other type of research.   On the other hand, if the purpose of the survey is to secure data to be used in improving or modifying

teaching procedures, the final object will be better served if the teachers themselves participate as much as possible in the various steps of the testing program. For example, if it is desired to discover whether too much or too little time is given to such subjects as arithmetic, spelling, reading, or history, the teacher will be able to fall into line with the new procedures more intelligently if she has become familiar with the method of determining the changes in procedure.

The director of the testing program will also need considerable assistance in having the tests scored. Several procedures are available. First, trained clerks may be employed. This method is advantageous in that it tends to minimize errors in scoring, and disadvantageous in that many schools have no provision in their budget for paying special clerks. Second, high school students may be used. If properly selected and trained, students can become very skillful and accurate in scoring. It is hardly justifiable, however, to take them away from their regular work to score test papers unless there is some compensating educational value in it, such as there might be for students in classes that are receiving normal school training preparatory to teaching. Third, teachers may be assigned to score the tests. This is perhaps the best procedure to follow, because the work will give teachers an insight into tests of a kind that will be valuable to them, particularly if the test has been given in their own grades. When the test has been given for such administrative or research purposes as we have suggested here, any undue personal interest in the results can be minimized by having each teacher score the test papers of another teacher's class. In case the testing program is to be followed by remedial steps in which the teacher is the chief factor, it would probably be better to let each teacher score the papers of her own pupils in order to gain as much knowledge and insight as possible through the examination of each paper.

When teachers are called on to do a large amount of scoring, as would be the case in a systematic general testing program, they should be relieved from other duties. Otherwise they will tend to dread or to resent the undertaking of a testing program. If no other method is feasible, the pupils should be dismissed from school for a half day or more while their teachers score the tests. It is often advantageous to have the teachers work in groups, which procedure makes it possible to organize scoring teams that greatly facilitate the scoring. It also makes it possible to explain any questions that arise in connection with the scoring and so secure greater uniformity. A third advantage is that the scoring can be checked to eliminate or minimize errors. Thus some scorers, if left to themselves, misinterpret certain directions for scoring, with the result that the particular error affects all the test papers scored by the teacher. In order to organize and control scoring teams properly, check the scoring, explain difficulties or misunderstandings, etc., the director or capable assistants should be present.

Testing for and by the teacher. In the foregoing discussion we have been concerned with tests that may be given for the information or use of administrative officers, supervisory officers, and teachers. Sometimes the interests of one group and sometimes those of another will predominate, according to the specific purposes for which the tests are given. However, in certain types of tests, such as diagnostic tests and remedial exercises, the classroom teacher is the person primarily concerned. She should therefore carry out all the steps in their administration. She may, of course, look to administrative and supervisory officers for help in the selection of the tests as well as in their administration. But in order to get the greatest possible educational results from their use, she should be thoroughly familiar with every phase of the application of the tests. This is well illustrated in

connection with the Compass Survey Tests, the Compass Diagnostic Tests, and the Economy Remedial Exercise Cards discussed in Chapter VII. These tests are intended to form an integral part of the teaching of arithmetic and the studying of the progress of each pupil. No one can take the teacher's place in performing these functions. In the administration of this type of test the teacher should be responsible for giving, scoring, and interpreting it, just as she is responsible for any other phase of classroom teaching. Since this type of test largely replaces the informal tests, the work required will not be excessive.

The costs of a testing program. While tests, such as the Stanford Achievement Test, may be obtained that range in cost from less than one cent per pupil to about seven or eight cents per pupil for a test battery, covering several subjects, many schools hesitate to venture on a testing program because of the financial outlay required. It may easily be shown that this attitude is one of false economy. In connection with a single item, that of retardation, it can be shown that a better placement of pupils will result in greater financial economy in running the schools. Thus the writer [1] has shown that during a given year six schools in the state of Idaho gave special promotion to 179 pupils after a testing program. During this year the average cost per pupil was $74.40. The special promotion referred to would therefore result in a saving of $13,317.60, to say nothing of the educational gain to the pupils. When it is recalled that retardation in many schools is as high as 40 to 50 per cent of the total enrollment, it can readily be understood that this condition will tend to result in congestion of pupils in some grades with the consequent need of either increasing the teaching load for each teacher or employing more teachers.

---

[1] I. N. Madsen, "Procedures Following a Testing Program." in *School and Society,* Vol. XIV, pages 600–605; December 24, 1921.

The most important gains from a properly planned testing program are, however, those resulting in the improvement of teaching rather than in the saving of money.

Test results in relation to parents and pupils. Especially when given for the first time in a community, standardized tests tend to arouse great interest and curiosity. This situation affords an opportunity to enlist public interest in the schools. The problem is to present the results of the program in the right way. One method of publicity has been described in Section I, which deals with the use of intelligence tests (page 225). The principles underlying that method are sound. Very little should be said to either parents or pupils about intelligence test results, and then only to those specifically concerned. Information is not withheld for the purpose of concealment, but because there are so many technical considerations involved that it would be hopeless to attempt to educate everyone to an adequate conception of the meaning of tests and test scores. In order to help parents and pupils to use intelligence test results for educational and vocational guidance, it is well to discuss with them the limitations and strengths of the pupil concerned. The situation should not be presented in a way that will suggest invidious comparisons. The presentation is the task of the individual best qualified; this usually is the director of the testing program, but may sometimes be the teacher. At present most teachers are themselves too much at sea as to the meaning of intelligence tests to undertake such explanations. They should therefore refer inquiries to the director.

With standardized achievement tests the situation is entirely different. Although they are better and more accurate measuring devices, they have the same meaning as ordinary teachers' examinations and so can be easily understood by both parents and pupils. Their results may therefore be used freely to inform both parents and pupils as to

the standing of the pupils concerned. The standing of other pupils should not be discussed, however, with a given parent or pupil. Such discussion starts gossip and may do considerable harm. It is sometimes useful to discuss in class the results obtained from an achievement test by the class as a whole, as many pupils may have encountered the same difficulties and time can be economized by a discussion of common errors. This is especially true of diagnostic tests. Such tests may show that a whole class is deficient in certain skills or information and that remedial exercises for the class as a whole are in order.

Testing in small schools. In the foregoing, testing programs for the larger schools have been considered. In the smaller schools each teacher must necessarily be more independent and self-reliant in the administration and use of tests. This is unfortunate for schools that have weak teachers. These schools, however, are no more unfortunate in regard to testing than in regard to the other functions that teachers perform. To the resourceful and self-reliant teacher the small school offers a challenge and an opportunity for self-development.

V. CRITERIA FOR THE SELECTION OF STANDARDIZED TESTS

Unfortunately the title of a standardized test does not necessarily reveal the purpose which that test will best serve. It is therefore necessary to set up criteria to aid in the evaluation of the various tests that are available for the testing of pupils in a given trait or subject. Even when such criteria are set up, considerable subjectivity enters into the amount of weight given by the appraiser to the various characteristics on the basis of which he selects a test. However, practice should result in more and more expertness, much as the use of score cards increases accuracy in judging stock. Some of the general characteristics of a good test were dis-

SCALE FOR RATING TESTS	NAMES OF TESTS					
Manual (5)						
Validity (15)						
Reliability (10)						
Reputation (5)						
Ease of Administration (Total 15) (a) Preparation (4)						
(b) Time limits (4)						
(c) Explanation needed (3)						
(d) Alternative forms (4)						
Ease of Scoring (Total 15) (a) Objectivity (10)						
(b) Time required (3)						
(c) Simplicity (2)						
Ease of Interpretation (Total 15) (a) Norms (5)						
(b) Directions for interpreting (4)						
(c) Class record (1)						
(d) Application of results (5)						
Convenient Packages (5)						
Typography and Makeup (5)						
Test Service (10)						
Total (100)						

FIG. 17. This scale is taken from *Test Service Bulletin, No. 13*, prepared by Arthur S. Otis and published by the World Book Company. Directions for the use of the scale are given in the bulletin.

cussed in Chapter II. The meaning and importance of validity and reliability were discussed in Chapter IV and in the chapters dealing with intelligence and achievement tests. Other important criteria for selecting tests are given in the following pages.

Objectivity of scoring. We have already seen in Chapter II that one important reason for the unsatisfactory results obtained in using the ordinary teacher's examination was the lack of objectivity of scoring. It is obvious that unless a test can be objectively scored its validity and reliability are seriously threatened. A good test should therefore provide scoring keys and complete directions for their use. We have seen in Chapter II also that even when these helps are provided, many scorers still make serious mistakes. Therefore in a testing program in which teachers are to do the scoring it is important to be sure that the scoring procedure is adequately explained to them and that adequate provision is made for it.

Administrative considerations. Validity, reliability, and objectivity, which have already been discussed, are qualities of primary importance in selecting standardized tests. Other criteria may be grouped together under the general term " administrative considerations." Other things being equal, these administrative considerations should also be given weight in the selection of tests for use in a school system. The more important of these elements of a test which should be of especial concern to the administrative authorities in a school system may be listed as follows :

1. Cost
2. Ease of administration
   a. Clearness and completeness of manual of directions
   b. Time limits for pupils in taking the test
   c. Number and equivalence of forms

d. Amount of training and preparation necessary for the examiner

e. Amount of time required for scoring

f. The kind of norms provided and other suggestions which may be used in interpreting the results

g. Provisions or suggestions made for pupils' and class records

h. Suggestions as to frequency and time of testing

A scale for rating tests. Dr. Arthur S. Otis has devised a rating scale, shown on page 262, by means of which the merits of several tests can be compared. The scale provides for rating each test on a number of characteristics, most of which have been discussed in this or other chapters. Complete directions accompany the scale, explaining how to use it. Such a scale helps the test user to keep in mind the characteristics that a good test should have and to decide whether a given test is suitable for a certain purpose or which of several tests should be used for a particular purpose.

References

ALEXANDER, CARTER. "Presenting Educational Measurements so as to Influence the Public Favorably." *Journal of Educational Research,* Vol. III (May, 1921), pages 345–358.

ARTHUR, GRACE. "A Quantitative Study of the Results of Grouping First-Grade Classes According to Mental Age." *Journal of Educational Research,* Vol. XII (October, 1925), pages 173–185.

BAGLEY, W. C. *Determinism in Education.* Warwick and York, Inc., Baltimore; 1925.

—— "Educational Determinism; or Democracy and the IQ." *School and Society,* Vol. XV (April 8, 1922), pages 373–384.

BERRY, CHARLES. "Classification by Tests of Intelligence of Ten Thousand First-Grade Pupils." *Journal of Educational Research,* Vol. VI (October, 1922), pages 185–203.

BROOKS, S. S. *Improving Schools by Standardized Tests.* Houghton Mifflin Company, Boston; 1922.

BUCKINGHAM, B. R. "The Public School Teacher as a Research Worker." *Journal of Educational Research,* Vol. XI (April, 1925), pages 235–243.

CHAPMAN, J. CROSBY. "The Unreliability of the Difference between Intelligence and Educational Ratings." *Journal of Educational Psychology,* Vol. XIV (February, 1923), pages 103–108.

COBB, MARGARET V. "The Limits Set to Educational Achievement by Limited Intelligence." *Journal of Educational Psychology,* Vol. XIII (November, 1922), pages 449–464.

COXE, WARREN W. "School Variation in General Intelligence." *Journal of Educational Research,* Vol. IV (October, 1921), pages 187–194.

DOLL, EDGAR A. "A Special Class Catechism." *Journal of Educational Research,* Vol. XII (October, 1925), pages 186–203.

GOODENOUGH, FLORENCE. "Efficiency in Learning and the Accomplishment Ratio." *Journal of Educational Research,* Vol. XII (November, 1925), pages 297–300.

HAWLEY, W. E. "The Effect of Clear Objectives on the Teaching of Reading." *Journal of Educational Research,* Vol. III (April, 1921), pages 254–260.

HERRING, JOHN P. "The Reliability of Accomplishment Differences." *Journal of Educational Psychology,* Vol. XV (November, 1924), pages 530–538.

HOLLINGWORTH, LETA. *Gifted Children, Their Nature and Nurture.* The Macmillan Company, New York; 1926.

INSKEEP, ANNIE DOLMAN. *Teaching Dull and Retarded Children.* The Macmillan Company, New York; 1926.

KALLOM, ARTHUR. "Intelligence Tests and the Classroom Teacher." *Journal of Educational Research,* Vol. V (May, 1922), pages 389–399.

KEANEY, JULIA F. "Teaching and Following up Supernormal Children in a Small Public School." *Journal of Educational Research,* Vol. VII (February, 1923), pages 145–148.

KEENER, E. E. "Results of Homogeneous Classification of Junior High School Pupils." *Journal of Educational Research,* Vol. XIV (June, 1926), pages 14–20.

—— "The Use of Measurements in a Small City School." *Journal of Educational Research,* Vol. III (March, 1921), pages 201–206.

McLEOD, BEATRICE, and IRVING, HELEN. "Objective Tests in the Rural Schools of Wyoming." *Journal of Educational Research,* Vol. XVII (January, 1928), pages 45–49.

MARTIN, A. LEILA, and PECHSTEIN, L. A. "Educational Tests for Retarded School Children." *Journal of Educational Research,* Vol. IX (May, 1924), pages 403–410.

MEAD, A. R. "Suggestions for the Training of Teachers in the Use of Educational Measurements." *Educational Administration and Supervision,* Vol. XII (January, 1926), pages 23–43.

MURDOCK, KATHERINE. "The Accomplishment Quotient — Finding and Using It." *Teachers College Record,* Vol. XXIII (May, 1922), pages 229–239.

NEWLON, JESSE H. "What Research Can Do for the Superintendent." *Journal of Educational Research,* Vol. VIII (September, 1923), pages 106–112.

PINTNER, RUDOLF, and NOBLE, HELEN. "The Classification of School Children According to Mental Age." *Journal of Educational Research,* Vol. II (November, 1920), pages 713–728.

POPENOE, HERBERT. "A Report of Certain Significant Deficiencies of the Accomplishment Quotient." *Journal of Educational Research,* Vol. XVI (June, 1927), pages 40–47.

SAAM, THEODORE. "Intelligence Testing as an Aid to Supervision." *Elementary School Journal,* Vol. XX (September, 1919), pages 26–32.

SEASHORE, C. E. *A Survey of Musical Talent in the Public Schools* (Studies in Child Welfare: First Series, No. 37: Vol. I, No. 2). University of Iowa, Iowa City; November, 1920.

STEDMAN, LULU M. *Education of Gifted Children.* World Book Company, Yonkers-on-Hudson, New York; 1924.

SYMONDS, PERCIVAL. "The Accuracy of Certain Standard Tests for School Classification." *Journal of Educational Research,* Vol. IX (April, 1924), pages 315–330.

—— "The Accuracy of Certain Standard Tests for School Sectioning and Marking." *Journal of Educational Psychology,* Vol. XV (October, 1924), pages 423–432.

TERMAN, L. M. "The Possibilities and Limitations of Training." *Journal of Educational Research,* Vol. X (December, 1924), pages 335–343.

—— "The Psychological Determinist; or Democracy and the IQ." *Journal of Educational Research,* Vol. VI (June, 1922), pages 57–62.

—— "The Use of Intelligence Tests in the Grading of School Children." *Journal of Educational Research,* Vol. I (January, 1920), pages 20–32.

THORNDIKE, E. L., et al. "Standard Tests and Their Use — A Symposium." *Teachers College Record,* Vol. XXVI (October, 1924), pages 93–116.

TOOPS, HERBERT A., and SYMONDS, P. M. "What Shall We Expect of the IQ?" *Journal of Educational Psychology,* Vol. XIII (December, 1922), pages 513–528, and Vol. XIV (January, 1923), pages 27–37.

TORGERSON, T. L. "Is Classification by Mental Ages and Intelligence Quotients Worth While?" *Journal of Educational Research,* Vol. XIII (March, 1926), pages 171–180.

VOGEL, MABEL, and WASHBURNE, CARLETON. "A Year of Winnetka Research." *Journal of Educational Research,* Vol. XVII (February, 1928), pages 90–101.

WASHBURNE, CARLETON. "The Individual System in Winnetka." *Elementary School Journal,* Vol. XXI (September, 1920), pages 52–68.

—— and RATHS, LOUIS. "The Selection of Under-Age Children for Entrance into School." *Educational Administration and Supervision,* Vol. XIV (March, 1928), pages 185–188.

WILSON, W. R. "The Misleading Accomplishment Quotient." *Journal of Educational Research*, Vol. XVII (January, 1928), pages 1–10.

WOODROW, HERBERT. "Mental Unevenness and Brightness." *Journal of Educational Psychology*, Vol. XIX (May, 1928), pages 289–302.

WOODY, CLIFFORD. "The Values of Educational Research to the Classroom Teacher." *Journal of Educational Research*, Vol. XVI (October, 1927), pages 172–178.

CHAPTER ELEVEN

The Improvement of Teachers' Examinations

The inflexibility of standardized tests. We noted in Chapter II the inaccuracy of school marks as obtained from the usual type of teachers' examinations. In the present chapter we shall discuss methods that may be followed to improve teachers' examinations. With such improvement, teachers' examinations can be made to serve many useful purposes that are difficult to serve by means of standardized tests. In the first place standardized tests are in their very nature inflexible. That is, they are prepared for a specific purpose and do not necessarily fit the exact need for testing on all occasions. Suppose that a teacher has just completed a given topic in history, geography, or arithmetic, and desires to test her pupils in that topic. It is not likely that she will be able to find a test that exactly covers the topic as it has been covered by the class. Again, suppose that she desires to give monthly tests to determine progress. It is not likely that she will be able to find standardized tests for every subject that will adequately meet the requirements for testing the pupils on the subject matter they have studied.

General purposes of teachers' examinations. We may summarize the main purposes of teachers' examinations in the elementary grades as follows:

1. To provide motivation or incentive to study.
2. To furnish a means of measurement in order to:
    a. Determine individual progress and make diagnosis.
    b. Determine promotion or graduation.
    c. Aid in sectioning or classifying pupils.
    d. Aid in educational and vocational guidance.
    e. Aid in the placement of new pupils.
    f. Award prizes, honors, etc.

3. To provide practice in organization and expression.

It is obvious that only valid and reliable tests are useful for these purposes. Therefore many teachers now attempt to apply the methods used in constructing standardized achievement tests to the construction of informal examinations for their own use.

Many other teachers have, however, objected to the use of objective tests, both standardized and informal, on the ground that they do not adequately provide practice in organization and expression. These teachers contend that written essay-type examinations are necessary in order to organize properly the subject matter which has just been studied and in order to provide valuable training in written English. There is little or no evidence to support these contentions. On the contrary, certain considerations lead us to think that, under actual conditions, the old-type written examinations quite often accomplish the exact opposite of these purposes.

In regard to the first assumption, we may refer to the principle that learning is specific; in other words, one learns precisely what one learns. Thus, learning to organize information and knowledge in written English does not necessarily result in the ability to organize such information or knowledge, orally or mentally, in the solution of problems. From the psychological standpoint, the best way to attain such abilities is to practice on them directly. From the practical standpoint, ability to organize information mentally and orally is likely to be more valuable to the pupil than the ability to make a formal parade of his knowledge in writing.

In regard to the second assumption, we may say that as a rule neither teachers nor pupils have in mind the organization and language-training function of such examinations. The pupils usually write under pressure, trying to finish in the

time allowed; consequently organization, spelling, punctuation, handwriting, etc., are neglected. Under typical conditions there is also much temptation to intellectual dishonesty. Questions often lend themselves to more than one interpretation, and pupils are tempted to select the interpretation that best fits the material that they happen to remember. Another common practice is the attempt to cover ignorance by mere verbosity. Moreover, the teacher often can give little or no attention to the organization or language functions of such tests. It is her chief concern to score the papers and to record the marks, usually under a time pressure. The line of least resistance is to score on the basis of the prima facie purpose of the test, which is to test the acquisition of knowledge or information in the subject concerned. Even when a teacher does point out their errors in organization and language to the pupils in a systematic way, she is likely to find that this distracts and confuses them with regard to errors in the subject itself. In the light of these considerations it is evident that the typical written examination is quite as likely to give the wrong as the right kind of training in organization and language. When in addition we recall the unreliability, as pointed out in Chapter II, of school marks that are based on the ordinary written examination, it is clear that this type of examination must give way to more objective methods.

Types of objective examinations.[1] Two main types of objective examinations, the recall and the recognition, together with some of the more common varieties of each type, were discussed in Chapter II. Concrete examples of these variations as well as others were given in Chapters VI, VII, and VIII in the discussion of intelligence and achievement tests.

---

[1] For more illustrative types of objective tests and for a number of selected complete examinations of the objective type see: G. M. Ruch, *The Objective or New-Type Examination*, Chapters VIII and IX. Scott, Foresman & Co., Chicago; 1929.

The present chapter will consider these different types from the standpoint of the teacher's own construction and use of informal objective examinations.

Simple recall tests. The following are examples of the simple recall statement in which the pupils answer by filling the blank space in each statement:

1. Columbus discovered America in the year   ....
2. Jamestown was founded during the year   ....
3. The name of the Indian princess who saved the life of Captain John Smith was   ....
4. During the year 1620 the Pilgrims founded the colony of   ....
5. The name of the ship in which the Pilgrims came to America was   ....

It will be noted that the scoring can be made very objective by preparing a key with the correct answers. Care must be taken to make sure that there is only one correct answer for each statement. A danger in constructing exercises of this type is that unimportant factual material will be included simply because it lends itself readily to this purpose. It is objected that this type of exercise seems to place an unwarranted emphasis on memorizing factual material. This is not a serious objection, however, if the exercises are constructed from important and carefully selected facts. In many life situations facts must be recalled in just this way if they are recalled at all.

The completion exercise. The completion exercise is a modification of the simple recall type. Instead of one blank to be filled by the pupil, there may be two or more. The following are examples of this type of exercise:

1. Columbus discovered .... in the year ....
2. Jamestown was founded during the year .... by ....
3. The name of the Indian princess who saved the life of .... was ....
4. During the year .... the Pilgrims founded the colony ̀of ....

5. The name of the ship in which the Pilgrims came to America and founded .... was ....

It will be observed that these exercises are based on the same material as the simple recall exercises. It happens that in each of the exercises above there are but two blanks to be filled, but it will readily be seen that it is not necessary to limit the number of blanks to two. However, when there are many blanks to be filled, there is danger of introducing a puzzle element. The resulting score would then be determined in part by a pupil's information and in part by his ability to solve language puzzles of this type. This in turn might lower the validity of the test. Another danger to be guarded against is ambiguity. For example, suppose the first statement to read: "America was discovered by —— during the year ——." The first blank could be correctly filled in more than one way. "Columbus," "The Northmen," "Leif Ericson," would all be correct. The date to be supplied in the second blank will depend on how the first blank is filled. The degree of ambiguity that is allowed to creep into this type of exercise increases the element of subjectivity in the scoring, because the scorer is then called upon to exercise his judgment as to the correctness of a given response.

True-false tests. The procedure in true-false tests is to devise statements some of which are true and some of which are false. An equal number of true and false statements are then arranged in a chance order. The following are examples of this type of exercise:

1. Columbus discovered America in the year 1492.   *True*   *False*
2. Jamestown was founded during the year 1632.   *True*   *False*
3. The name of the Indian princess who saved the life of Captain John Smith was Pocahontas.   *True*   *False*
4. During the year 1620 the Pilgrims founded the colony of Plymouth.   *True*   *False*
5. The name of the ship in which the Pilgrims came to America was the Ironsides.   *True*   *False*

In the directions the pupils may be told to guess or not to guess. The scoring can be made entirely objective by using prepared scoring keys. The most common procedure to determine a pupil's score is to subtract the number of items wrongly marked from the number correctly marked. This amounts to counting off two points for each item incorrectly marked, and is designed to offset the element of chance.

The true-false type of exercise is one of the most popular because the statements are comparatively easy to make and the scoring can be made entirely objective. The time allowed to the pupils permits the introduction of as many as a hundred or more true-false statements for a single test, thus enabling the teacher to cover the whole field quite thoroughly. In this type of exercise, however, as in the foregoing types, it is important to guard against ambiguity. The statements should be clearly either true or false. It is important also to guard against including unimportant material or material that is too easy or too difficult, since such material lowers the reliability and validity of the test. Beginners should exercise care also in scoring this type of test. It is the writer's experience that the right-minus-wrong method of scoring is confusing to many inexperienced scorers and even to some experienced ones. Types of errors likely to occur in scoring this test are shown in Chapter II, page 30.

An objection rather frequently made to the use of true-false exercises is that putting false statements before pupils is likely to have bad educational consequences. Very little experimental evidence on this point is available. Roberts and Ruch,[1] however, point out that on purely theoretical grounds one might argue with equal truth that there are positive teaching effects in true statements. These authors

[1] Hazel M. Roberts and G. M. Ruch, "The Negative Suggestion Effect of True-False Tests," in *Journal of Educational Research,* Vol. XVIII, pages 112–116; September, 1928.

also point out that the effects of negative suggestion by false statements must be permanent in order to be of serious consequences. In reporting an experiment on this matter these investigators conclude that negative suggestion effects exist, at least for short intervals; that the amount is smaller than was generally believed on theoretical grounds; that " the *net* suggestion effect of a true-false test appears to be *positive*, not negative "; and that all effects, both positive and negative, appear to be transitory.

The multiple-choice test. As the name implies, the multiple-choice type of exercise offers a choice of more than one response. The pupils are directed to underline the word or item that makes the statement true. The following are examples of this type of exercise:

1. Columbus discovered America in the year —
     1368   1492   1504   1584
2. Jamestown was founded during the year —
     1536   1564   1607   1620
3. The name of the Indian princess who saved the life of Captain John Smith was —
     Sacajawea   Minnehaha   Pocahontas   Red Wing
4. During the year 1620 the Pilgrims founded the colony of —
     Boston   New York   Philadelphia   Plymouth
5. The name of the ship in which the Pilgrims came to America was —
     **The Flying Cloud   The Mayflower   Old Ironsides
     The Lusitania**

It is, of course, obvious that the number of choices may be more or less than the four used in the above illustrations. If the number of choices is less than four, it is customary to make a correction for the element of chance in scoring. Thus if the test is of the alternative-response type, the scoring procedure is the same as for the true-false type. If there are three choices, a common method is to subtract one-half the number wrong from the number right. If there are four or more choices, it is not usually considered necessary to correct for the chance element. The multiple-choice method is

applicable to many kinds of material. By requiring the pupil to make a selection from among several possible responses, this method tends to require thinking and judgment as well as mere memory.

Best-answer exercises. The best-answer type of exercise is really a variation of the multiple-choice type. A common method offers a choice of three statements in answer to a given proposition, as follows:

1. Columbus made his first voyage across the Atlantic because —
   he was trying to find a new route to India.
   he was fleeing from the Spanish king.
   he was pursuing the enemy.
2. The Pilgrims came to America because —
   they wanted religious freedom.
   they believed they would find much gold.
   they were seeking adventure.

The pupils are usually directed to place a check mark beside, or to indicate in some other given way, the statement that is the best answer. There is no reason why there should not be more than three choices or why there should be only one acceptable answer. By making the exercise more complex in this way, thinking and judgment are drawn upon to a greater extent. This type of exercise is scored in the same manner as the multiple-choice tests.

The matching exercise. There are many different methods of arranging test material in the form of matching exercises. One method is illustrated in the following:

	MEN	EVENTS
( )	Washington	1. British general
( )	Jefferson	2. Confederate general
( )	Lincoln	3. Invented steamboat
( )	Robert E. Lee	4. First President of the United States
( )	Robert Fulton	5. Louisiana Purchase
( )	Cornwallis	6. Emancipation of slaves

In connection with this type of exercise the pupils may be directed to write in the parentheses opposite the name of each man the number indicated for the event or descriptive term with which he is associated.   Thus the number " 4 " would be written in the parentheses opposite the name of Washington.   It is evident that if the items are carefully chosen such exercises can be made entirely objective.   The exercise may be made more difficult by including more items in one group than in the other — which makes it difficult for the pupil to do the matching by a process of elimination.

Need of care in the choice of type and in the construction of test items.   The foregoing illustrations of objective types of examinations include by no means all the possible types or variations of types.   However, they do represent those more commonly used.   The illustrations are taken from one subject — United States history.   They might equally well have been taken from geography, nature study, civics, English, physiology, or other subjects.   In preparing objective tests for any of these subjects, it will be found that one or more of the foregoing types can be used.   In all cases care should be taken to select the types that are best suited to the material on which the pupils are to be tested.   There is also need for care in selecting and wording each item of a test.   The mere use of objective forms for testing does not guarantee the validity of the test.   Doubtless there are many teachers who will remain careless in the construction of this type of test just as they are in the construction of the written essay-type examination questions.   There are teachers who make up a list of examination questions extemporaneously.   It is evident that this procedure gives little opportunity for either thoughtful selection of questions or care in their construction. In this respect the objective examination has a distinct advantage over the essay-type examination.   In its very nature it requires that some time and thought be given to the selec-

tion and statement of the various items. Some teachers mistakenly begrudge this extra amount of time. As a matter of fact, the time and labor spent in the construction of objective tests is much more than regained in the greater speed and facility with which the scoring may be done.

Common errors in constructing objective tests. Among the more common mistakes made in constructing objective tests are the five suggested here. (1) Trivial or unimportant matter may be included. It is obvious that such material serves merely as " filler " and contributes nothing to testing a pupil's knowledge of the important phases of the subject. (2) Items that are too easy or too hard are sometimes included. The objection to such items is similar to the foregoing. For example, items so easy that every pupil answers them do not aid in differentiating between pupils as to achievement. If a test were made up entirely of such items, every pupil would make a perfect score. Thus the teacher would be no better able to distinguish the progress of the different pupils or to make diagnosis of their individual needs than she was before giving the test. Similarly, items so difficult that few or no pupils can answer them would contribute little or nothing to the discovery of differences in achievement. (3) Ambiguous items are sometimes included. It is clear that if a question can be interpreted and answered in more than one way it introduces an element of subjectivity into the scoring that will tend to lower both the validity and the reliability of the test. (4) Items taken from subject matter that has received little or no attention in study or recitation are sometimes used. Such items tend to creep into a test because they can easily be taken from a tabulation in the textbook, although they are important only in connection with a minor point. (5) Occasionally too few items are used. Table 49 (page 279) indicates that 100 items give reliability coefficients ranging from .71 to .90. The number

of items included in tests given during a term should therefore be well above this number.

Experimental studies of different types of tests.[1] Questions often arise as to which kinds of tests are the least time-consuming so far as the pupils are concerned, which kinds are most valid and reliable, whether the nature of the directions to the pupils makes any difference, etc. A number of studies in answer to these questions have already been made. While final conclusions have not been arrived at as yet, some important information has been obtained. Thus Ruch [2] worked out experimentally the following time schedule for the type of objective tests that are indicated for elementary-grade pupils:

Recall type	3 to 5 items per minute
Recognition type	4 to 6 items per minute
True-false type	5 to 8 items per minute

Ruch states that these schedules are approximate and that they depend upon the subject matter and the degree of difficulty of the items. It is also well to recall that pupils will show wide differences in the rate at which they work. Each teacher should be able to work out fairly accurate time schedules for the different types of tests that she uses in her teaching.

Ruch also gives reliability coefficients for five different types of tests that " were suitable in difficulty for twelfth-grade pupils." These are shown in Table 49.

In a more recent study Ruch and Stoddard [3] publish valid-

---

[1] For detailed studies of comparative validities, reliabilities, working times, and difficulties of various types of objective test items see: G. M. Ruch, *The Objective or New-Type Examination,* Chapter XI. Scott, Foresman & Co., Chicago; 1929.

[2] G. M. Ruch, *Improvement of the Written Examination,* page 97. Scott, Foresman & Co., Chicago; 1924.

[3] G. M. Ruch and G. D. Stoddard, *Tests and Measurements in High School Instruction,* pages 289–291. World Book Company, Yonkers-on-Hudson, New York; 1927.

TABLE 49

RELIABILITY COEFFICIENTS OF THE FIVE TYPES OF EXAMINATIONS [1]

TYPE	FORM A vs. FORM B (i.e., 50 items vs. 50 items)	RELIABILITY OF 100 ITEMS (by Brown's) Formula	N
Recall . . .	.81 ± .010	.90	562
5-response . .	.80 ± .021	.89	137
3-response . .	.60 ± .037	.75	134
2-response . .	.74 ± .027	.85	135
True-false . .	.56 ± .040	.71	133

ty coefficients obtained in Grades VII, VIII, XI, and XII or true-false, 2-response, 3-response, 5-response, and 7-reponse tests, which range from .675 to .926. This study hows validity coefficients for each of these tests when pupils ire instructed to guess and also when they are instructed not o guess. The tests are also scored by two different methods; 1amely, without correction for chance and with correction or chance. In their conclusions from their studies these 1uthors state:

For practical purposes the differences in the various techniques ire not large enough to rule out any particular combination of nethods completely. Both the validity and reliability can be raised easily by the expedient of longer tests. It is probably true hat 150 items under "guess" instructions and without correction ire at least as valid as 100 items given under "do not guess" direc-ions and with corrections for chance.

Conversion of point scores into grades. The most direct nethod for converting point scores into grades is to use the 1ormal frequency curve. On the basis of this curve many schools have adopted a letter-grade system as follows:

6 per cent of pupils should receive a grade of A.
25   "   "   "   "   "   "   "   " B.

[1] G. M. Ruch, *Improvement of the Written Examination*, page 111. Scott, Foresman & Co., Chicago; 1924.

38 per cent of pupils should receive a grade of C.
25 " " " " " " " " D.
6 " " " " " " " " F.

Obviously more or fewer than five letter grades may be used and different per cents than those given above may be adopted. In Chapter III it was pointed out that the normal frequency curve can be derived by expanding a binomial. The distribution above is suggested by the expansion of $(x + y)^4$, which yields $x^4 + 4\,x^3y + 6\,x^2y^2 + 4\,xy^3 + y^4$. The sum of the coefficients of the five terms is 16, and dividing the coefficients of each term by this number yields approximately the per cents given above. With this distribution in mind it becomes an easy matter to convert point scores into grades by assigning a grade of " A " to the 6 per cent of pupils obtaining the highest score, a grade of " B " to the next highest 25 per cent of pupils, etc. It should be kept in mind, however, that the normal curve operates best when the group is fairly large. Thus it should not be adhered to slavishly with small groups. It is justifiable to assign more grades either above or below the middle grade of " C " if there is evidence that the class is not a normal group.

There are several methods by means of which a teacher can determine whether her class is normal in ability. One method is to give a group test of intelligence from which both the average intelligence and the distribution for the group can be determined. Similar information can be obtained by giving an achievement test. Another method is available in cases where the teacher has given her own objective test to more than one group of pupils who have studied the same subject under the same conditions. This method may also be used to establish norms from which the point scores may be converted into grades. Such norms are especially valuable to determine grades for pupils in small classes where the normal curve may not operate.

A real difficulty arises when the pupils of a grade are divided into several sections on the basis of their ability to achieve. Suppose, for example, a grade is divided in this way into three sections. It is then evident that if scores (expressed as subject grades) are distributed on the basis of the normal curve, they will not have the same meaning for the different sections. A grade of " A " in the superior section would not be the same as a grade of " A " in either of the other two sections. Similarly for other letter grades. One method proposed for eliminating this difficulty is to give the same test to all three sections and then distribute the grades on the basis of the normal curve without regard to section. However, several objections to this procedure immediately arise. In the first place, the superior section would have a monopoly on the " A " and " B " grades, while the inferior section would soon become discouraged by the fact that it obtained the majority of " F " and " D " grades. There is another objection; namely, that often where this method of classification of pupils is used the course of study is not the same for the different groups. Thus a test suited to the superior section would not necessarily be suited to the other sections. Another method, which attempts to take account of these facts, proposes to add a subscript to each letter grade to indicate the section in which it has been earned. Thus a grade of " A " given to a pupil would be written as " A_1" in the superior section, as " A_2" in the average section, and as " A_3" in the inferior section.

In using the normal curve as a basis for distributing grades, a problem arises in connection with the interpretation that should be given to the lowest grade assigned, which in our illustration is " F." It is a debatable question whether any failing grades should be given in the elementary schools. It can be argued, for example, that in these grades pupils attend school under compulsory school laws and that they have very

little option as to what they study; that if they do as well as their native endowment permits, it is a species of cruelty to assign failing grades, and such a practice may defeat some functions of the school by developing " inferiority complexes," antisocial attitudes, etc.   It can be argued further that the situation is not analogous to that in colleges or professional schools where the interests of the public and of future patients or clients must also be considered.   What may be good practice in these schools is not, therefore, necessarily good practice in the elementary schools.   However, this is a matter to be determined in connection with the general policy of the school, and if it is decided that no failing grade shall be given, the lowest grade may still be used to designate a lower quality of work than that represented by the higher grades.[1]

Another common question in regard to school marks is whether such matters as effort, attitude, deportment, attendance, etc., should be included in determining a pupil's grade. This question should be answered in the negative.   A pupil's grade should be based entirely upon his actual scholastic achievement in his class.   This is not because the other factors mentioned are unimportant, but because the inclusion of anything but actual achievement in a school mark tends to obscure its meaning.   For example, different teachers would not be influenced in like degree by various extraneous factors.   Thus, in evaluating the school marks given by different teachers in a subject like arithmetic, there would be no way of telling how much of a given mark represented actual achievement in arithmetic, how much represented effort, how much deportment, how much honesty, etc.   If pupils are to be rated in these matters, ratings apart from the school marks should be used.

---

[1] Failing grades, of course, can be avoided if pupils are grouped on the basis of ability as illustrated by the Trinidad plan, or by the use of some method of individual instruction as illustrated by the Winnetka plan.

Advantages and limitations of teachers' objective examinations. Among the advantages of objective examinations the following may be mentioned: (1) The use of the normal curve as a basis for distributing school marks tends to make the marks given by different teachers more comparable. It is obvious that the teacher who gives 50 per cent of " A's " to her pupils places a different value on " A " than is placed by a teacher who gives but 5 per cent of " A's." In some cases entire school systems may err in giving too high a per cent of high grades. Pupils from such a school who move to a school that pays more attention to the normal curve are quite often overestimated by their new teachers, unless it is known that there is a tendency in the first school to grade too high. In such a case the marks may be discounted. (2) Objective examinations require less of the pupil's time and less of the teacher's time and yet make it possible to test the pupil more thoroughly than do essay-type examinations. (3) Objective tests are more reliable and valid than essay-type examinations, or they can be made so with a reasonable amount of effort.

Among the limitations of objective examinations we have already called attention to the fact that it seems to be difficult to use these tests in such subjects as arithmetic and English composition. However, the difficulties encountered in these subjects may lessen as other methods are devised. In the meantime certain phases of these subjects are susceptible to objective testing. Thus arithmetic exercises may be scored on the basis of " right answer." Similarly, certain types of errors in English composition, such as misspellings, errors in punctuation, errors in capitalization, etc., can be determined objectively. In the case of English composition, rating scales may be used to evaluate the " story " or "literary " value of each pupil's composition. Also, in grading problems in arithmetic, greater objectivity will result if a set

of rules is carefully prepared. If there are several teachers of this subject in the same grade, greater uniformity in grading will be secured if they will make out these rules together and come to an understanding before grading as to how they shall apply them.

Measurement of the " intangibles." It has always been considered an important function of the school to cultivate in pupils such traits as appreciation of art, music, and literature, to develop character and morality, and to inculcate good citizenship and patriotism. This belief has persisted in spite of the fact that the inculcation of these traits is usually sought as a by-product of the more formal functions, such as the development of skill or knowledge in music, art, literature, history, etc. They are also sought as by-products of extracurricular activities, or even as a result of the influence of the character and personality of the teacher.

Many educators have held that these traits are too subtle and intangible for objective measurement. It is indeed true that but little progress has been made in the direction of measuring them. Among the reasons for this lack of progress we may note the following : First, there seems to be an absence in pupils of observable or unambiguous reactions that can be measured objectively. For example, what are the reactions of a pupil that may be used to measure the extent of his artistic or musical appreciation? We cannot depend upon his own statement, for this is too subjective. We cannot judge very accurately by such reactions as facial expressions or spontaneous exclamations, for some pupils are much more demonstrative than others. In the measurement of moral traits some progress has been made by psychologists, but the methods in use cannot readily be adapted to objective testing executed by the classroom teacher. In the case of patriotism and citizenship we are dealing with functions that do not have full expression until after school days are

over. It would, therefore, be necessary to discover and to measure a pupil's reactions at the time of his measurement, which would predict his later reactions to be manifested in the expression of citizenship and patriotism. Informational testing will not suffice, for knowledge of what is right does not necessarily result in doing what is right, particularly when activity is to take place many years in the future.

Another difficulty is that authorities do not agree on the meaning of some of these traits; that is, on their definition. *Patriotism* affords a good example of such a trait. To what extent does patriotism consist of the glorification of national military heroes, of victorious wars, of pacifists, of captains of industry, of leaders in science, of great artists, of great preachers, of literary men, and so forth? There is violent disagreement about the selection of elements that should be glorified and about the extent to which such glorification indicates patriotism.

Until our educational philosophers and sociologists succeed in agreeing more nearly on defining in concrete terms our objectives in the development of these traits, and until our experts in educational measurements succeed in devising techniques for their measurement, we shall in all probability continue to have the present unsatisfactory situation. To the extent, however, that such traits as these are associated with the more accepted and measurable functions of education, or are by-products of them, we may improve our measurement of the former by improving our measurement of the latter.

References

ASKER, WILLIAM. "The Reliability of Tests Requiring Alternative Responses." *Journal of Educational Research,* Vol. IX (March, 1924), pages 234–240.

FOSTER, R. R., and RUCH, G. M. "On Corrections for Chance in Multiple-Response Tests." *Journal of Educational Psychology,* Vol. XVIII (January, 1927), pages 48–51.

FRITZ, MARTIN F. "Guessing in a True-False Test." *Journal of Educational Psychology*, Vol. XVIII (November, 1927), pages 558–562.

HAHN, H. H. "A Criticism of Tests Requiring Alternative Responses." *Journal of Educational Research*, Vol. VI (October, 1922), pages 236–240.

HENMON, V. A. C. "Limitations of Educational Tests." *Journal of Educational Research*, Vol. VII (March, 1923), pages 185–198.

MONROE, W. S. *Written Examinations and Their Improvement* (University of Illinois Bulletin, Vol. XX, No. 7; and Bureau of Educational Research Bulletin, No. 9). University of Illinois, Urbana; 1922.

—— and SOUDERS, LLOYD B. *The Present Status of Written Examinations and Suggestions for Their Improvement* (University of Illinois Bulletin, Vol. XXI, No. 13; and Bureau of Educational Research Bulletin, No. 17). University of Illinois, Urbana; November 26, 1923.

ODELL, C. W. "Another Criticism of Tests Requiring Alternative Responses." *Journal of Educational Research*, Vol. VII (April, 1923), pages 326–330.

—— *Traditional Examinations and New-Type Tests.* The Century Company, New York; 1928.

ORLEANS, JACOB S., and SEALY, GLENN A. *Objective Tests.* World Book Company, Yonkers-on-Hudson, New York; 1928.

PATERSON, D. G. *Preparation and Use of New-Type Examinations.* World Book Company, Yonkers-on-Hudson, New York; 1925.

—— and LANGLIE, T. A. "Empirical Data on the Scoring of True-False Tests." *Journal of Applied Psychology*, Vol. IX (1925), pages 339–348.

RUCH, G. M. *Improvement of the Written Examination.* Scott, Foresman & Co., Chicago; 1924.

—— *The Objective or New-Type Examination.* Scott, Foresman & Co., Chicago; 1929.

—— and DEGRAFF, M. H. "Corrections for Chance and 'Guess' vs. 'Do not Guess' Instruction in Multiple-Response Tests." *Journal of Educational Psychology*, Vol. XVII (September, 1926), pages 368–375.

—— and STODDARD, G. D. "Comparative Reliabilities of Five Types of Objective Examinations." *Journal of Educational Psychology*, Vol. XVI (February, 1925), pages 89–103.

RUSSELL, CHARLES. *Classroom Tests.* Ginn & Co., New York; 1926.

WEST, PAUL. "A Critical Study of the Right Minus Wrong Method." *Journal of Educational Research*, Vol. VIII (June, 1923), pages 1–9.

WORCESTER, D. A. "Prevalent Errors in New-Type Examinations." *Journal of Educational Research*, Vol. XVIII (June, 1928), pages 48–52.

APPENDIX

The Calculation of the Coefficient of Correlation

WHILE there are several formulas for computing the coefficient of correlation, the following form of the Pearson method is usually found to be the most convenient:[1]

$$r = \frac{\dfrac{\Sigma xy}{N} - c_x c_y}{\sigma_x \sigma_y}$$

In this formula r is the symbol for coefficient of correlation. The other symbols have either been explained earlier (page 67) or will be explained in the solution of an illustrative problem (page 289). When a relationship such as that shown in Table 24 (page 72) is solved by this formula, it results in a coefficient of correlation of $+ 1.00$. On the other hand, Table 26 (page 74) would yield a coefficient of correlation of $- 1.00$. All coefficients of correlation lie between these two limits. In actual practice neither perfect positive nor perfect negative correlation is ever obtained in computing the correlation between grades in two school subjects or between scores in two standardized tests. Most of them are positive correlations and lie between 0.3 and 0.9.

We shall now proceed to show how this formula may be applied in computing the coefficient of correlation between the history and arithmetic scores obtained in testing ninety-two seventh-grade pupils. Table 50 shows the results tabulated in the form of a double distribution table.

By referring back to the formula for computing the coefficient of correlation, it will be seen that all of the symbols with

---

[1] For the derivation of this formula see some standard text on statistical methods, such as T. L. Kelley's *Statistical Method*. (Published by the Macmillan Company, New York; 1923.)

TABLE 50

Showing Distribution of History and Arithmetic Scores Obtained in Testing Ninety-Two Seventh-Grade Pupils (x = Arithmetic Scores; y = History Scores)

History Scores	Arithmetic Scores							f_y	d_y	fd_y	$fd_y{}^2$
	20–39	40–59	60–79	80–99	100–119	120–139	140–159				
170–179							1	1	7	7	49
160–169								0	6	0	0
150–159						1		1	5	5	25
140–149					1	1		2	4	8	32
130–139					2	3		5	3	15	45
120–129				2	7			9	2	18	36
110–119				4	9	1		14	1	14	14
100–109				3	8			11	0		
90–99			1	11	5	1		18	— 1	— 18	18
80–89			4	5	4			13	— 2	— 26	52
70–79		1	2	5	1	1		10	— 3	— 30	90
60–69		1	2	2				5	— 4	— 20	80
50–59		1						1	— 5	— 5	25
40–49	1		1					2	— 6	— 12	72
f_x	1	3	10	32	37	8	1	92		— 44	538
d_x	— 3	— 2	— 1	0	1	2	3				
fd_x	— 3	— 6	— 10		37	16	3	37			
$fd_x{}^2$	9	12	10		37	32	9	109			

the exception of Σxy appear in computing the standard deviation (σ), as illustrated in Table 22. We may therefore proceed to determine the values for these symbols for the two frequency distributions which appear in Table 50. The four columns at the right of the table give the necessary data for the variable y (history scores) and the four rows at the foot

of the table give similar data for the variable x (arithmetic scores). Solving for these values, we obtain:

$$N = 92 \text{ (number of cases for both distributions)}$$
$$c_x = \tfrac{37}{92} = .40$$
$$c_y = \frac{-44}{92} = -.48$$
$$\sigma_x^2 = \tfrac{109}{92} - .16 = 1.02$$
$$\sigma_y^2 = \tfrac{538}{92} - .23 = 5.62$$
$$\sigma_x = 1.01$$
$$\sigma_y = 2.37$$
$$\Sigma xy = 1.39 \text{ (obtained as follows)}$$

Each frequency in the correlation table is multiplied by the product of its deviations from both assumed means. Thus in Table 50 there is a frequency of 1 opposite the interval in Column I for history scores 170–179 and below the interval for arithmetic scores 140–159. The deviation of this frequency in terms of class interval from the assumed mean of the history distribution is 7, and from the assumed mean of the arithmetic distribution it is 3. Thus the product in this case is 21. In obtaining this product, the signs of the deviations must be regarded. For this reason the products obtained from the frequencies in the upper right and from the lower left sections of the correlation table will be positive, while those obtained from the upper left and the lower right sections will be negative. Σxy is the algebraic sum of these products. We may now proceed to substitute in the formula given on page 287 the values listed in the preceding paragraph. Following this procedure we obtain:

$$r = \frac{\tfrac{139}{92} - (.40 \times -.48)}{1.01 \times 2.37}$$
$$= \frac{1.51 + .19}{2.39} = \frac{1.70}{2.39}$$
$$= .71$$

When the coefficient of correlation has been calculated, it is desirable also to calculate its probable error, which in general has the same significance in relation to the coefficient of correlation that the probable errors discussed in Chapter IV have to the respective group measures for which they are calculated. The formula for the probable error of the coefficient of correlation (P.E.$_r$) is written as follows:

$$\text{P.E.}_r = .6745 \, \frac{1 - r^2}{\sqrt{N}}$$

Illustrating with the data from Table 50 in which r is .71 and N is 92, we obtain:

$$\text{P.E.}_r = .6745 \, \frac{1 - .5041}{\sqrt{92}} = \frac{.3345}{\sqrt{92}} = .035$$

P.E.$_r$ is usually written with the coefficient of correlation for which it is calculated. Thus the results obtained above would be written:

$$r = .71 \pm .035$$

We can now say that the chances are even that the true coefficient of correlation lies between .675 and .745. We can also say that the chances are about four to one that the true correlation lies between .64 and .78. It follows from the foregoing that the trustworthiness of a given coefficient of correlation depends, in part, upon the size of its probable error.

INDEX

Ability grouping, pros and cons, 227 ff.

Accomplishment quotient, 249 ff.; formula for, 250; limitations of, 251

Achievement tests, 137 ff.; selected list of, 199 ff.; uses of, 237

Anderson, W. N., 170, 172

Arithmetic tests, standardized, 137 ff.; practice, 147; relation of, to courses of study, 150

Army Alpha Test, 113; interpretation of scores, 115; use in schools and colleges, 118 ff.

Army Beta Test, 117; description of, 117 ff.; interpretation of, 120

Ashbaugh, E. J., 17, 172

Ayres Handwriting Scale, 161 ff.

Ayres, Leonard P., 2, 161, 170

Ayres Spelling Scale, 171 ff.

Best-answer exercises, 275

Binet, Alfred, 1

Binet-Simon Test, 96; Stanford Revision of, 97; summary of Stanford Revision, 97 ff.; learning to use Stanford Revision, 105; Herring Revision, 106; Kuhlmann Revision, 107

Buckingham, B. R., 172, 250

Buckingham-Stevenson Place Geography Tests, 181

Burgess Picture Supplement Scale, 154

Burk, F. L., 224

Cambridge plan, 224

Cattell, J. McKeen, 1

Central tendency, measures of, 51; inadequacy of measure of, 64 ff.

Charters Diagnostic Language Tests, 177, 180

Class interval, 40

Classification, use of both intelligence and achievement tests for, 229

Coin tossing, 43 ff.

Compass Diagnostic Tests, 145 ff., 241

Compass Survey Tests, 144, 145

Completion exercise, 271–272

Corning, Hobart M., 225, 226, 227

Correlation, 70; meaning of, 70 ff.; positive correlation illustrated, 72; negative correlation illustrated, 74; calculation of coefficient of, 73, 74, 287 ff.; uses of, 75 ff.; probable error of, 290

Courtis Practice Tests, 148

Courtis Supervisory Tests in Geography, 181

Crabbs, L. M., 205

Dalton plan, 224

Deciles, 63

Dickson, V. E., 33, 104, 106

Distribution, skewed, 44 ff.; bi-modal and multi-modal, 47 ff.

Economy Remedial Exercise Cards, 149

Eldridge, R. C., 170

English, 176 ff.

Fernald, Grace, 229, 240

Franzen, Raymond H., 32, 249, 252, 253, 254

Freeman, F. N., 162, 251

Frequency curve, normal, 43

Frequency polygon, 42

Frequency surface, 41 ff.

Frequency table, 38 ff.

Galton, Sir Francis, 1, 88

Gates, A. I., 205, 206, 245

Gaussian curve, 43

Geography, 181 ff.